Bootstrap confidence sets under model misspecification

Dissertation

zur Erlangung des akademischen Grades
Dr. Rer. Nat.
im Fach Mathematik

eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakultät
Humboldt-Universität zu Berlin

von
Dipl.-Math. Mayya Zhilova

Präsident der Humboldt-Universität zu Berlin:
Prof. Dr. Jan-Hendrik Olbertz

Dekan der Mathematisch-Naturwissenschaftlichen Fakultät:
Prof. Dr. Elmar Kulke

Gutachter:

1. Prof. Dr. Vladimir Spokoiny
2. Prof. Dr. Gilles Blanchard
3. Prof. Dr. Victor Chernozhukov

eingereicht am: 30. Juni 2015
Tag der Verteidigung: 02. November 2015

To my parents

Abstract

The thesis studies a multiplier bootstrap procedure for construction of likelihood-based confidence sets in two cases. The first one focuses on a single parametric model, while the second case extends the construction to simultaneous confidence estimation for a collection of parametric models. Theoretical results justify the validity of the bootstrap procedure for a limited sample size n, a large number of considered parametric models K, growing parameters' dimensions, and possible misspecification of the parametric assumptions.

In the case of one parametric model the bootstrap approximation works if p^3/n is small, where p is the parameter's dimension. The main result about bootstrap validity continues to apply even if the underlying parametric model is misspecified under the so-called Small Modelling Bias condition. If the true model deviates significantly from the considered parametric family, the bootstrap procedure is still applicable but it becomes a bit conservative: the size of the constructed confidence sets is increased by the modelling bias.

For the problem of construction of simultaneous confidence sets we suggest a multiplier bootstrap procedure for estimating the quantiles of the joint distribution of the likelihood ratio statistics, and for adjustment of the confidence level for multiplicity. Theoretical results state the bootstrap validity taking into account the bootstrap correction for multiplicity, they require the quantity $(\log K)^{12} p_{\max}^3/n$ to be small, where $p_{\max}$ is the maximal parameter dimension. Here we also consider the situation when the parametric models are misspecified. If the models' misspecification is significant, then the bootstrap critical values exceed the true ones and the simultaneous bootstrap confidence set becomes conservative.

The theoretical approach is based on several approximating bounds: non-asymptotic square-root Wilks theorem, Gaussian approximation of Euclidean norm of a sum of independent vectors, comparison and anti-concentration bounds for Euclidean norms of Gaussian vectors. Numerical experiments for misspecified linear, logistic, local constant and local quadratic regressions nicely confirm our theoretical results.

Zusammenfassung

Diese Arbeit befasst sich mit einem Multiplier-Bootstrap Verfahren für die Konstruktion von Likelihood-basierten Konfidenzbereichen in zwei verschiedenen Fällen. Im ersten Fall betrachten wir das Verfahren für ein einzelnes parametrisches Modell und im zweiten Fall erweitern wir die Methode, um Konfidenzbereiche für eine ganze Familie von parametrischen Modellen simultan zu schätzen.

Theoretische Resultate zeigen die Validität der Bootstrap-Prozedur für eine potenziell begrenzte Anzahl an Beobachtungen n, eine große Anzahl K an betrachteten parametrischen Modellen, wachsende Parameterdimensionen und eine mögliche Misspezifizierung der parametrischen Annahmen. Im Falle eines einzelnen parametrischen Modells funktioniert die Bootstrap-Approximation, wenn p^3/n klein ist, wobei p hier die Parameterdimension bezeichnet. Das Hauptresultat über die Validität des Bootstrap gilt unter der sogenannten Small-Modelling-Bias Bedingung auch im Falle, dass das parametrische Modell misspezifiert ist. Wenn das wahre Modell signifikant von der betrachteten parametrischen Familie abweicht, ist das Bootstrap Verfahren weiterhin anwendbar, aber es führt zu etwas konservativeren Schätzungen: die Konfidenzbereiche werden durch den Modellfehler vergrößert.

Für die Konstruktion von simultanen Konfidenzbereichen entwickeln wir ein Multiplier-Bootstrap Verfahren um die Quantile der gemeinsamen Verteilung der Likelihood-Quotienten zu schätzen und eine Multiplizitätskorrektur der Konfidenzlevels vorzunehmen. Theoretische Ergebnisse zeigen die Validität des Verfahrens unter der Bedingung das $(\log K)^{12} p_{\max}^3/n$ klein ist, wobei $p_{\max}$ die maximale Parameterdimension bezeichnet. Hier betrachten wir auch wieder den Fall, dass die parametrischen Modelle misspezifiziert sind. Wenn die Misspezifikation signifikant ist, werden Bootstrap-generierten kritischen Werte größer als die wahren Werte sein und die Bootstrap-Konfidenzmengen sind konservativ.

Die theoretische Untersuchung basiert auf einer Reihe von Approximationsresultaten: dem nicht-asymptotischen square-root Wilks Theorem, der Gaußschen Approximation der euklidischen Norm einer Summe von unabhängigen Vektoren, Vergleichsresultate und Antikonzentrationsschranken für Normen von Gaußschen Vektoren. Numerische Experimente für misspezifizierte lineare, logistische, lokal konstante und lokal quadratische Regression bestätigen unsere theoretischen Ergebnisse.

Acknowledgements

First and foremost, I would like thank my supervisor Vladimir Spokoiny for his careful guidance and generous support during my PhD studies. I am very grateful to him for many inspiring and fruitful discussions, for his insightful comments and for being actively interested in my work during the whole period of my studies. I would also like to thank Prof. Reiß and Prof. Härdle for their important comments on my talks, which helped to improve this work.

I am very thankful to my friends and colleagues at the Weierstrass Institute for lots of helpful discussions, for their support and many cheerful moments. I am especially grateful to Andreas Andresen, Christian Bayer, Natalia Bochkina, Thorsten Dickhaus, Kirill Efimov, Roland Hildebrand, Marcel Ladkau, Nina Loginova, Hilmar Mai, Peter Mathé, Mario Maurelli, Tigran Nagapetyan, Jörg Polzehl, Konstantin Schildknecht, John Schoenmakers, Jens Stange, Karsten Tabelow, Niklas Willrich and Jianing Zhang.

An excellent working environment of the Weierstrass Institute helped me a lot to do this research, and I am cordially grateful to the institute's directorate and administration for that. This research was supported by the German Research Foundation (DFG) through the Collaborative Research Center 649 "Economic Risk".

Contents

Chapter 1

Introduction

Bootstrap is a technique for making statistical inference about unknown population by resampling from an observed data set. The bootstrap was firstly introduced by Efron (1979), and since then became one of the most powerful and common tools in statistical confidence estimation and hypothesis testing. Bootstrap procedure is particularly useful for making inference in complicated statistical models, since it leads to a bootstrap world (see Efron and Tibshirani (1994), pp. 86-88), where all the objects are available for computation. Many versions and extensions of the original bootstrap method have been proposed in the literature; see e.g. Wu (1986); Mammen (1993); Newton and Raftery (1994); Janssen (1994); Barbe and Bertail (1995); Shao and Tu (1995); Horowitz (2001); Chatterjee and Bose (2005); Ma and Kosorok (2005); Chen and Pouzo (2009); Lavergne and Patilea (2013); Bücher and Dette (2013); Chen and Pouzo (2015) among many others.

This work focuses on the multiplier bootstrap procedure which attracted a lot of attention last time due to its nice theoretical properties and numerical performance. We mention the papers Chatterjee and Bose (2005), Arlot et al. (2010a,b) and Chernozhukov et al. (2013a, 2014a,b) for the most advanced recent results. Chatterjee and Bose (2005) showed some results on asymptotic bootstrap consistency in a very general framework: for estimators obtained by solving estimating equations. Arlot et al. (2010a) constructed a non-asymptotical confidence bound in ℓ_s-norm ($s \in [1, \infty]$) for the mean of a sample of high dimensional i.i.d. Gaussian vectors (or with a symmetric and bounded distribution), using generalized bootstrap for resampling of the quantiles. Arlot et al. (2010b) extended that results for the multiple testing problems for mean values of coordinates of high-dimensional i.i.d. Gaussian vectors with unknown covariance matrix. They provided non-asymptotic control for the family-wise error rate using resampling-type procedures. Chernozhukov et al. (2013a) presented a number

of non-asymptotic results on Gaussian approximation and multiplier bootstrap for maxima of sums of high-dimensional vectors (with a dimension possibly much larger than a sample size) in a very general set-up. As one of the applications the authors considered the problem of multiple hypothesis testing in the framework of approximate means. They derived non-asymptotic results for the general stepdown procedure by Romano and Wolf (2005) with improved error rates and in high-dimensional setting. Chernozhukov et al. (2014a) showed how this technique applies to the problem of constructing an honest confidence set in nonparametric density estimation. Chernozhukov et al. (2014b) extended the results from maxima to the class of sparsely convex sets.

The present work makes a further step in studying the multiplier bootstrap method in the problems of confidence estimation and simultaneous confidence estimation by a quasi maximum likelihood method. For a rather general parametric model, we consider likelihood-based confidence sets (and simultaneous confidence sets) with the radius determined by a multiplier bootstrap. The aim of the study is to check the validity of the bootstrap procedure in the following setting:

1. the sample size n is fixed;

2. the parametric models can be misspecified;

3. the dimensions of the considered parametric models can be dependent on the sample size n;

4. in the case of simultaneous confidence estimation the number K of the parametric models can be exponentially large w.r.t. n.

In the case of a single parametric model our results explicitly describe the error term of the bootstrap approximation. This particularly allows to track the impact of the parameter dimension p, the sample size n in the quality of the bootstrap procedure. As one of the corollaries, we show bootstrap validity under the constraint "p^3/n-small". Chatterjee and Bose (2005) stated results under the condition "p/n-small" but their results only apply to low dimensional projections of the MLE vector. In the likelihood based approach, the construction involves the Euclidean norm of the MLE, which leads to completely different tools and results. Chernozhukov et al. (2013a) allowed a huge parameter dimension with "$\log(p)/n$ small" but they essentially work with a family of univariate tests which again differs essentially from the maximum likelihood approach.

Another interesting and important issue is the impact of the model misspecification on the accuracy of bootstrap approximation. A surprising corollary of our error bounds

is that the bootstrap confidence set can be used even if the underlying parametric model is slightly misspecified under the so-called *small modelling bias* **(SmB)** condition. If the modelling bias becomes large, the bootstrap confidence sets are still applicable, but they become more and more conservative. The **(SmB)** condition is given in Section 2.5 and it is consistent with classical bias-variance relation in nonparametric estimation. Numerical experiments in Section 2.4 nicely confirm our theoretical results. Below in this chapter we describe the problem and the theoretical approach in more detail, see Sections 1.1-1.3.

The problem of simultaneous confidence estimation appears in numerous practical applications when a confidence statement has to be made simultaneously for a collection of objects, e.g. in safety analysis in clinical trials, gene expression analysis, population biology, functional magnetic resonance imaging and many others. See e.g. Miller (1981); Westfall (1993); Manly (2006); Benjamini (2010); Dickhaus (2014), and references therein. This problem is also closely related to construction of simultaneous confidence bands in curve estimation, which goes back to Working and Hotelling (1929). For an extensive literature review about constructing the simultaneous confidence bands we refer to Hall and Horowitz (2013), Liu (2010), and Wasserman (2006).

A simultaneous confidence set requires a probability bound to be constructed jointly for several possibly dependent statistics. Therefore, the critical values of the corresponding statistics should be chosen in such a way that the joint probability distribution achieves a required family-wise confidence level. This choice can be made by multiplicity correction of the marginal confidence levels. The Bonferroni correction method (Bonferroni (1936)) uses a probability union bound, the corrected marginal significance levels are taken equal to the total level divided by the number of models. This procedure can be very conservative if the considered statistics are positively correlated and if their number is large. The Šidák correction method (Šidák (1967)) is more powerful than Bonferroni correction, however, it also becomes conservative in the case of large number of dependent statistics.

Most of the existing results about simultaneous bootstrap confidence sets and resampling-based multiple testing are asymptotic (with sample size tending to infinity), see e.g. Beran (1988, 1990); Hall and Pittelkow (1990); Härdle and Marron (1991); Shao and Tu (1995); Hall and Horowitz (2013), and Westfall (1993); Dickhaus (2014). The results based on asymptotic distribution of maximum of an approximating Gaussian process (see Bickel and Rosenblatt (1973); Johnston (1982); Härdle (1989)) require a huge sample size n, since they yield a coverage probability error of order $(\log(n))^{-1}$ (see Hall (1991)). Some papers considered an alternative approach in context of

confidence band estimation based on the approximation of the underlying empirical processes by its bootstrap counterpart. In particular, Hall (1993) showed that such an approach leads to a significant improvement of the error rate (see also Neumann and Polzehl (1998); Claeskens and Van Keilegom (2003)). Chernozhukov et al. (2014a) constructed honest confidence bands for nonparametric density estimators without requiring the existence of limit distribution of the supremum of the studentized empirical process: instead, they used an approximation between sup-norms of an empirical and Gaussian processes, and anti-concentration property of suprema of Gaussian processes.

In many modern applications the sample size cannot be large, and/or it can be smaller than a parameter dimension, for example, in genomics, brain imaging, spatial epidemiology and microarray data analysis, see Leek and Storey (2008); Kim and van de Wiel (2008); Arlot et al. (2010b); Cao and Kosorok (2011), and references therein. For the recent results on resampling-based simultaneous confidence sets in high-dimensional finite sample set-up we refer to the papers by Arlot et al. (2010b) and Chernozhukov et al. (2013a, 2014a,b), cited above in this section.

The present work's set-up 1-4, in contrast with the paper by Chernozhukov et al. (2014b), does not require the sparsity condition, in particular the dimensions $p_1, \ldots, p_K$ of each parametric family may grow with the sample size. Moreover, the simultaneous likelihood-based confidence sets are not necessarily convex, and the parametric assumption can be violated.

The considered simultaneous multiplier bootstrap procedure involves two main steps: estimation of the quantile functions of the likelihood ratio statistics, and multiplicity correction of the marginal confidence level. Theoretical results state the bootstrap validity in the setting 1-4, taking in account the multiplicity correction. The resulting approximation bound requires the quantity $(\log K)^{12} p_{\max}^3 / n$ to be small. The log-factor here is suboptimal and can probably be improved. We particularly address the problem of the impact of the model misspecification. For the problem of simultaneous confidence estimation we introduce the "simultaneous small modelling bias condition" $\widehat{\mathbf{(SmB)}}$ given in Section 3.4.2. This condition roughly means that all the parametric models are close to the true distribution. If $\widehat{\mathbf{(SmB)}}$ condition is fulfilled, then the bootstrap approximation is accurate, otherwise the simultaneous bootstrap confidence set is still applicable, however, it becomes conservative. This property is nicely confirmed by the numerical experiments in Section 3.3.

Sections 1.1 - 1.3 below provide an introduction to the case of a single parametric model and give an overview of the theoretical approach. The problem of constructing

the simultaneous confidence sets is described in Section 1.4.

1.1 Likelihood-based confidence sets

The idea of constructing confidence intervals using the likelihood function goes back to Fisher (see Fisher (1956); Hudson (1971) and references therein). The Wilks phenomenon described below justifies this idea.

Let the data sample $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ consist of *independent* random observations and belong to the probability space $(\Omega, \mathcal{F}, I\!P)$. We do not assume that the observations Y_i are identically distributed, moreover, no specific parametric structure of $I\!P$ is being required. In order to explain the idea of the approach we start here with a parametric case, however the assumption (1.1) below is not required for the results. Consider some known parametric family $\{I\!P(\boldsymbol{\theta})\} \stackrel{\text{def}}{=} \{I\!P(\boldsymbol{\theta}) \ll \mu_0, \boldsymbol{\theta} \in \Theta \subset I\!R^p\}$. If $I\!P \in \{I\!P(\boldsymbol{\theta})\}$, then the true parameter $\boldsymbol{\theta}^* \in \Theta$ is such that

$$I\!P \equiv I\!P(\boldsymbol{\theta}^*) \in \{I\!P(\boldsymbol{\theta})\}, \tag{1.1}$$

and the initial problem of finding the properties of unknown distribution $I\!P$ is reduced to the equivalent problem for the finite-dimensional parameter $\boldsymbol{\theta}^*$. The parametric family $\{I\!P(\boldsymbol{\theta})\}$ induces the log-likelihood process $L(\boldsymbol{\theta})$ of the sample $\boldsymbol{Y}$:

$$L(\boldsymbol{\theta}) = L(\boldsymbol{Y}, \boldsymbol{\theta}) \stackrel{\text{def}}{=} \log\left(\frac{dI\!P(\boldsymbol{\theta})}{d\mu_0}(\boldsymbol{Y})\right)$$

and the maximum likelihood estimate (MLE) of $\boldsymbol{\theta}^*$:

$$\widetilde{\boldsymbol{\theta}} \stackrel{\text{def}}{=} \operatorname{argmax}_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}). \tag{1.2}$$

The asymptotic Wilks phenomenon Wilks (1938) states that for the case of i.i.d. observations with the sample size tending to the infinity the likelihood ratio statistic converges in distribution to $\chi^2_p/2$, where p is the parameter dimension:

$$2\big\{L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*)\big\} \xrightarrow{w} \chi^2_p, \ n \to \infty.$$

Define the likelihood-based confidence set as

$$\mathcal{E}(\mathfrak{z}) \stackrel{\text{def}}{=} \Big\{\boldsymbol{\theta} : L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}) \le \mathfrak{z}^2/2\Big\},$$

then the Wilks phenomenon implies

$$I\!P\Big\{\boldsymbol{\theta}^* \in \mathcal{E}(\mathfrak{z}_{\chi^2_p}(\alpha))\Big\} \to \alpha, \ n \to \infty,$$

where $\mathfrak{z}^2_{\chi^2_p}(\alpha)$ is the $(1-\alpha)$-quantile for the χ^2_p distribution. This result is very important and useful under the parametric assumption, i.e. when (1.1) holds. In this case the limit distribution of the likelihood ratio is independent of the model parameters or in other words it is *pivotal*. By this result a sufficiently large sample size allows to construct the confidence sets for $\boldsymbol{\theta}^*$ with a given coverage probability. However, a possibly low speed of convergence of the likelihood ratio statistic makes the asymptotic Wilks result hardly applicable to the case of small or moderate samples. Moreover, the asymptotical pivotality breaks down if the parametric assumption (1.1) does not hold (see Huber (1967); White (1982)) and, therefore, the whole approach may be misleading if the model is considerably misspecified. If the assumption (1.1) does not hold, then the "true" parameter is defined by the projection of the true measure $I\!P$ on the parametric family $\{I\!P(\boldsymbol{\theta})\}$:

$$\boldsymbol{\theta}^* \stackrel{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\theta}\in\Theta} I\!E L(\boldsymbol{\theta}), \tag{1.3}$$

or equivalently

$$\boldsymbol{\theta}^* \stackrel{\text{def}}{=} \operatorname*{argmin}_{\boldsymbol{\theta}\in\Theta} \mathrm{KL}\left(I\!P, I\!P(\boldsymbol{\theta})\right).$$

The recent results by Spokoiny (2012a, 2013) provide a non-asymptotic version of square-root Wilks phenomenon for the case of misspecified model. It holds with an exponentially high probability

$$\left| \sqrt{2\{L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*)\}} - \|\boldsymbol{\xi}\| \right| \le \Delta_{\mathrm{W}} \simeq \frac{p}{\sqrt{n}}, \tag{1.4}$$

where $\boldsymbol{\xi} \stackrel{\text{def}}{=} D_0^{-1}\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}^*)$, $D_0^2 \stackrel{\text{def}}{=} -\nabla^2_{\boldsymbol{\theta}} I\!E L(\boldsymbol{\theta}^*)$. The bound is non-asymptotical, the approximation error term Δ_{W} has an explicit form (the precise statement is given in Theorem A.2, Section A.1), and it depends on the parameter dimension p, sample size n, and the probability of the random set on which the result holds.

Due to this bound, the original problem of finding a quantile of the LR test statistic $L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*)$ is reduced to a similar question for the approximating quantity $\|\boldsymbol{\xi}\|$. The difficulty here is that in general $\|\boldsymbol{\xi}\|$ is non-pivotal, it depends on the unknown distribution $I\!P$ and the target parameter $\boldsymbol{\theta}^*$.

1.2 Multiplier bootstrap procedure for the case of one parametric model

In the present work we study the *multiplier bootstrap* (or *weighted bootstrap*) procedure for estimation of the quantiles of the likelihood ratio statistic. The idea of the procedure

is to mimic a distribution of the likelihood ratio statistic by reweighing its summands with random multipliers independent of the data:

$$L^{\circ}(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \sum\nolimits_{i=1}^{n} \log\left(\frac{d\mathbb{P}_{\boldsymbol{\theta}}}{d\mu_0}(Y_i)\right) u_i.$$

Here the probability distribution is taken conditionally on the data $\boldsymbol{Y}$, which is denoted by the sign $^{\circ}$ (also $\mathbb{E}^{\circ}$ and Var° denote expectation and variance operators w.r.t. the probability measure conditional on $\boldsymbol{Y}$). The random weights $u_1, \ldots, u_n$ are i.i.d., independent of $\boldsymbol{Y}$ and it holds for them: $\mathbb{E}^{\circ}(u_i) = 1$, $\mathrm{Var}^{\circ}(u_i) = 1$, $\mathbb{E}^{\circ} \exp(u_i) < \infty$. Therefore, the multiplier bootstrap induces the probability space conditional on the data $\boldsymbol{Y}$. A simple but important observation is that $\mathbb{E}^{\circ} L^{\circ}(\boldsymbol{\theta}) \equiv L(\boldsymbol{\theta})$, and hence,

$$\mathrm{argmax}_{\boldsymbol{\theta}}\, \mathbb{E}^{\circ} L^{\circ}(\boldsymbol{\theta}) = \mathrm{argmax}_{\boldsymbol{\theta}}\, L(\boldsymbol{\theta}) = \widetilde{\boldsymbol{\theta}}. \tag{1.5}$$

This means that the target parameter in the bootstrap world is precisely known and it coincides with the maximum likelihood estimator $\widetilde{\boldsymbol{\theta}}$ conditioned on $\boldsymbol{Y}$, therefore, the bootstrap likelihood ratio statistic $L^{\circ}(\widetilde{\boldsymbol{\theta}}^{\circ}) - L^{\circ}(\widetilde{\boldsymbol{\theta}}) \stackrel{\text{def}}{=} \sup_{\boldsymbol{\theta} \in \Theta} L^{\circ}(\boldsymbol{\theta}) - L^{\circ}(\widetilde{\boldsymbol{\theta}})$ is fully computable and leads to a simple computational procedure for the approximation of the distribution of $L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*)$.

1.3 Theoretical justification of the bootstrap for the case of one parametric model

The goal of the present study is to show the validity of the described multiplier bootstrap procedure in a fixed sample size set-up, and to obtain an explicit bound on the error of coverage probability. In other words, we are interested in non-asymptotic approximation of the distribution of $\{L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*)\}^{1/2}$ with the distribution of $\{L^{\circ}(\widetilde{\boldsymbol{\theta}}^{\circ}) - L^{\circ}(\widetilde{\boldsymbol{\theta}})\}^{1/2}$. So far there exist very few theoretical non-asymptotic results about bootstrap validity. Classical asymptotic tools for showing the bootstrap consistency are based on weak convergence arguments which are not applicable in the finite sample set-up. Some different methods have to be applied. In particular, the approach of Liu (1988) based on Berry-Esseen theorem can be extended to a finite sample set-up with a univariate parameter. For a high dimensional parameter space, important contributions are done in the recent papers by Arlot et al. (2010a) and Chernozhukov et al. (2013a, 2014b). The latter papers used a Gaussian approximation, Gaussian comparison, and Gaussian anti-concentration technique in high dimension. Our approach is similar but we combine it with the square-root Wilks expansion and

use Pinsker's inequality for Gaussian comparison and anti-concentration steps. The main steps of our theoretical study are illustrated by the following scheme:

$$\begin{array}{lcccccc} & & \overset{\text{sq-Wilks theorem}}{} & & \overset{\text{Gauss. approx.}}{} & & \\ \boldsymbol{Y}\text{-world:} & \sqrt{2L(\widetilde{\boldsymbol{\theta}}) - 2L(\boldsymbol{\theta}^*)} & \underset{p/\sqrt{n}}{\approx} & \|\boldsymbol{\xi}\| & \underset{(p^3/n)^{1/8}}{\overset{w}{\approx}} & \|\overline{\boldsymbol{\xi}}\| & \\ & & & & & w \wr \sqrt{p}\delta^2_{\text{smb}} \quad \overset{\text{Gauss. compar.}}{} & \\ \overset{\text{Bootstrap world:}}{} & \sqrt{2L^\circ(\widetilde{\boldsymbol{\theta}}^\circ) - 2L^\circ(\widetilde{\boldsymbol{\theta}})} & \underset{p/\sqrt{n}}{\approx} & \|\boldsymbol{\xi}^\circ\| & \underset{(p^3/n)^{1/8}}{\overset{w}{\approx}} & \|\overline{\boldsymbol{\xi}}^\circ\| & \end{array} \tag{1.6}$$

where

$$\boldsymbol{\xi}^\circ \overset{\text{def}}{=} \boldsymbol{\xi}^\circ(\boldsymbol{\theta}^*) \overset{\text{def}}{=} D_0^{-1}\nabla_{\boldsymbol{\theta}}\left[L^\circ(\boldsymbol{\theta}^*) - I\!\!E^\circ L^\circ(\boldsymbol{\theta}^*)\right]. \tag{1.7}$$

The vectors $\overline{\boldsymbol{\xi}}$ and $\overline{\boldsymbol{\xi}}^\circ$ are zero mean Gaussian and they mimic the covariance structure of the vectors $\boldsymbol{\xi}$ and $\boldsymbol{\xi}^\circ$: $\overline{\boldsymbol{\xi}} \sim \mathcal{N}(0, \operatorname{Var}\boldsymbol{\xi})$, $\overline{\boldsymbol{\xi}}^\circ \sim \mathcal{N}\bigl(0, \operatorname{Var}^\circ \boldsymbol{\xi}^\circ\bigr)$.

The error term shown below each arrow corresponds to the i.i.d. case considered in details in Section A.3.1. The upper line of the scheme corresponds to the $\boldsymbol{Y}$-world, the lower line - to the bootstrap world. In both lines we apply two steps for approximating the corresponding likelihood ratio statistics. The first approximating step is the non-asymptotic square-root Wilks theorem: the bound (1.4) for the $\boldsymbol{Y}$-case and a similar statement for the bootstrap world, which is obtained in Theorem A.4, Section A.2. The corresponding error is of order $p/\sqrt{n}$ for the case of i.i.d. observations; in the bootstrap world the square-root Wilks expansion implies

$$\left|\sqrt{2L^\circ(\widetilde{\boldsymbol{\theta}}^\circ) - 2L^\circ(\widetilde{\boldsymbol{\theta}})} - \|\boldsymbol{\xi}^\circ(\widetilde{\boldsymbol{\theta}})\|\right| \le \mathtt{C}p/\sqrt{n} \tag{1.8}$$

for $\boldsymbol{\xi}^\circ(\boldsymbol{\theta}) \overset{\text{def}}{=} D_0^{-1}\nabla_{\boldsymbol{\theta}}\bigl[L^\circ(\boldsymbol{\theta}) - I\!\!E^\circ L^\circ(\boldsymbol{\theta})\bigr]$. In our approximation diagram we use $\boldsymbol{\xi}^\circ(\boldsymbol{\theta}^*)$ instead of $\boldsymbol{\xi}^\circ(\widetilde{\boldsymbol{\theta}})$ which is more convenient for the GAR step and is justified by Lemma A.2 showing that $\bigl\|\boldsymbol{\xi}^\circ(\widetilde{\boldsymbol{\theta}}) - \boldsymbol{\xi}^\circ(\boldsymbol{\theta}^*)\bigr\| \le \mathtt{C}p/\sqrt{n}$.

The next step is called *Gaussian approximation* (GAR) which means that the distribution of the Euclidean norm $\|\boldsymbol{\xi}\|$ of a centered random vector $\boldsymbol{\xi}$ is close to the distribution of the similar norm of a Gaussian vector $\|\overline{\boldsymbol{\xi}}\|$ with the same covariance matrix as $\boldsymbol{\xi}$. A similar statement holds for the vector $\boldsymbol{\xi}^\circ$. Thus, the initial problem of comparing the distributions of the likelihood ratio statistics is reduced to the comparison of the distributions of the Euclidean norms of two centered normal vectors $\overline{\boldsymbol{\xi}}$ and $\overline{\boldsymbol{\xi}}^\circ$ (Gaussian comparison). This last step links their distributions and encloses the approximating scheme. The Gaussian comparison step is done by computing the Kullback-Leibler divergence between two multivariate Gaussian distributions (i.e.

by comparison of the covariance matrices of $\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}^*)$ and $\nabla_{\boldsymbol{\theta}} L^{\circ}(\boldsymbol{\theta}^*)$) and applying Pinsker's inequality (Lemma B.5. At this point we need to introduce the "small modelling bias" condition **(SmB)** from Section 3.4.2. It is formulated in terms of the following nonnegative-definite $p \times p$ symmetric matrices:

$$H_0^2 \stackrel{\text{def}}{=} \sum\nolimits_{i=1}^{n} I\!\!E \left[\nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}^*) \nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}^*)^\top \right], \tag{1.9}$$

$$B_0^2 \stackrel{\text{def}}{=} \sum\nolimits_{i=1}^{n} I\!\!E \left[\nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}^*) \right] I\!\!E \left[\nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}^*) \right]^\top \tag{1.10}$$

for $\ell_i(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \log\left(\frac{d I\!\!P_{\boldsymbol{\theta}}}{d\mu_0}(Y_i) \right)$, so that $\operatorname{Var}\{\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}^*)\} = H_0^2 - B_0^2$. If the parametric assumption (1.1) is true or if the data $\boldsymbol{Y}$ are i.i.d., then it holds $I\!\!E[\nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}^*)] \equiv 0$ and $B_0^2 = 0$. The **(SmB)** condition roughly means that the bias term B_0^2 is small relative to H_0^2. Below we show that the Kullback-Leibler distance between the distributions of two Gaussian vectors $\overline{\boldsymbol{\xi}}$ and $\overline{\boldsymbol{\xi}}^{\circ}$ is bounded by $p\|H_0^{-1} B_0^2 H_0^{-1}\|^2/2$. The **(SmB)** condition precisely means that this quantity is small (in scheme (1.6) it is denoted by $\sqrt{p}\delta^2_{\text{smb}}$). In Section A.3.4 the value $\|H_0^{-1} B_0^2 H_0^{-1}\|$ is evaluated for some commonly used models: the case of i.i.d. observations, generalized linear model and linear quantile regression. Below we distinguish between two situations: when the condition **(SmB)** is fulfilled and the opposite case. Theorems 2.1 and 2.2 in Section 2.1 deal with the first case, it provide the cumulative error term for the coverage probability of the confidence set (1.1), taken at the $(1-\alpha)$-quantile computed with the multiplier bootstrap procedure. The proof of this result (see Section D.1.1) summarizes the steps of scheme (1.6). The biggest term in the full error is induced by Gaussian approximation and requires the ratio p^3/n to be small. In the case of a "large modelling bias", i.e. when **(SmB)** does not hold, the multiplier bootstrap procedure continues to apply. It turns out that the bootstrap quantiles increase with the growing modelling bias, hence, the confidence set based on it remains valid, however, it may become conservative. This result is given in Theorem 2.4 of Section 2.1. The problems of Gaussian approximation and comparison for the Euclidean norm are considered in Sections C.1 and B.4 in general terms independently of the statistical setting of the thesis, and might be interesting by themselves. Section B.4 presents also an anti-concentration inequality for the Euclidean norm of a Gaussian vector. This inequality shows how the deviation probability changes with a threshold. The general results on GAR are summarized in Theorem B.1 and restated in Proposition D.1 for the setting of scheme (1.6). These results are also non-asymptotic with explicit errors and apply under the condition that the ratio p^3/n is small.

In Theorem 2.3 we consider the case of a scalar parameter $p = 1$ with an improved error term. Furthermore in Section 2.3 we propose a modified version of a quantile

function based on a smoothed probability distribution. In this case the obtained error term is also better than in the general result.

1.4 Simultaneous confidence sets

Here we use the notations from Section 1.1. Let the random data

$$\boldsymbol{Y} \stackrel{\text{def}}{=} (Y_1, \ldots, Y_n)^\top \tag{1.11}$$

consist of *independent* observations Y_i, and belong to the probability space $(\Omega, \mathcal{F}, I\!P)$. The sample size n is *fixed*. $I\!P$ is an *unknown probability distribution* of the sample $\boldsymbol{Y}$. Consider K parametric families of probability distributions:

$$\{I\!P_k(\boldsymbol{\theta})\} \stackrel{\text{def}}{=} \{I\!P_k(\boldsymbol{\theta}) \ll \mu_0, \boldsymbol{\theta} \in \Theta_k \subset I\!R^{p_k}\}, \quad k = 1, \ldots, K.$$

Each parametric family induces the quasi log-likelihood function for $\boldsymbol{\theta} \in \Theta_k \subset I\!R^{p_k}$

$$\begin{aligned} L_k(\boldsymbol{Y}, \boldsymbol{\theta}) &\stackrel{\text{def}}{=} \log\left(\frac{dI\!P_k(\boldsymbol{\theta})}{d\mu_0}(\boldsymbol{Y})\right) \\ &= \sum\nolimits_{i=1}^{n} \log\left(\frac{dI\!P_k(\boldsymbol{\theta})}{d\mu_0}(Y_i)\right). \end{aligned} \tag{1.12}$$

It is important that we *do not require* that $I\!P$ belongs to any of the known parametric families $\{I\!P_k(\boldsymbol{\theta})\}$, that is why the term *quasi* log-likelihood is used here. Below in this section we consider two popular examples of simultaneous confidence sets in terms of the quasi log-likelihood functions (1.12). Namely, the simultaneous confidence band for local constant regression, and multiple quantiles regression.

The target of estimation for the misspecified log-likelihood $L_k(\boldsymbol{\theta})$ is such a parameter $\boldsymbol{\theta}_k^*$, that minimises the Kullback-Leibler distance between the unknown true measure $I\!P$ and the parametric family $\{I\!P_k(\boldsymbol{\theta})\}$:

$$\boldsymbol{\theta}_k^* \stackrel{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta_k} I\!E L_k(\boldsymbol{\theta}). \tag{1.13}$$

The maximum likelihood estimator is defined as:

$$\widetilde{\boldsymbol{\theta}}_k \stackrel{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta_k} L_k(\boldsymbol{\theta}).$$

The parametric sets Θ_k have dimensions p_k, therefore, $\widetilde{\boldsymbol{\theta}}_k, \boldsymbol{\theta}_k^* \in I\!R^{p_k}$. For $1 \le k, j \le K$ and $k \ne j$ the numbers p_k and p_j can be unequal.

The likelihood-based confidence set for the target parameter $\boldsymbol{\theta}_k^*$ is

$$\mathcal{E}_k(\mathfrak{z}) \stackrel{\text{def}}{=} \left\{\boldsymbol{\theta} \in \Theta_k : L_k(\widetilde{\boldsymbol{\theta}}_k) - L_k(\boldsymbol{\theta}) \le \mathfrak{z}^2/2\right\} \subset I\!R^{p_k}. \tag{1.14}$$

Let $\mathfrak{z}_k(\alpha)$ denote the $(1-\alpha)$-quantile of the corresponding square-root likelihood ratio statistic:

$$\mathfrak{z}_k(\alpha) \stackrel{\text{def}}{=} \inf\left\{\mathfrak{z} \geq 0 : I\!P\left(L_k(\widetilde{\boldsymbol{\theta}}_k) - L_k(\boldsymbol{\theta}_k^*) > \mathfrak{z}^2/2\right) \leq \alpha\right\}. \tag{1.15}$$

Together with (1.14) this implies for each $k = 1, \ldots, K$:

$$I\!P\Big(\boldsymbol{\theta}_k^* \in \mathcal{E}_k\left(\mathfrak{z}_k(\alpha)\right)\Big) \geq 1 - \alpha. \tag{1.16}$$

Thus $\mathcal{E}_k(\mathfrak{z})$ and the quantile function $\mathfrak{z}_k(\alpha)$ fully determine the marginal $(1-\alpha)$-confidence set. The simultaneous confidence set requires a correction for multiplicity. Let $\mathfrak{c}(\alpha)$ denote a maximal number $c \in (0, \alpha]$ s.t.

$$I\!P\left(\bigcup\nolimits_{k=1}^{K}\left\{\sqrt{2L_k(\widetilde{\boldsymbol{\theta}}_k) - 2L_k(\boldsymbol{\theta}_k^*)} > \mathfrak{z}_k(c)\right\}\right) \leq \alpha. \tag{1.17}$$

This is equivalent to

$$\mathfrak{c}(\alpha) \stackrel{\text{def}}{=} \sup\left\{c \in (0,\alpha] : I\!P\left(\max_{1\leq k\leq K}\left\{\sqrt{2L_k(\widetilde{\boldsymbol{\theta}}_k) - 2L_k(\boldsymbol{\theta}_k^*)} - \mathfrak{z}_k(c)\right\} > 0\right) \leq \alpha\right\}. \tag{1.18}$$

Therefore, taking the marginal confidence sets with the same confidence levels $1 - \mathfrak{c}(\alpha)$ yields the simultaneous confidence bound of the total level $1-\alpha$. The value $\mathfrak{c}(\alpha) \in (0, \alpha]$ is the correction for multiplicity. In order to construct the simultaneous confidence set using this correction, one has to estimate the values $\mathfrak{z}_k(\mathfrak{c}(\alpha))$ for all $k = 1, \ldots, K$. By definition this problem splits into two subproblems:

1. **Marginal step.** Estimation of the marginal quantile functions $\mathfrak{z}_1(\alpha)$, ..., $\mathfrak{z}_K(\alpha)$ given in (1.15).

2. **Correction for multiplicity.** Estimation of the correction for multiplicity $\mathfrak{c}(\alpha)$ given in (1.18).

If the 1-st problem is solved for any $\alpha \in (0,1)$, the 2-nd problem can be treated by calibrating the value α s.t. (1.18) holds. It is important to take into account the correlation between the likelihood ratio statistics $L_k(\widetilde{\boldsymbol{\theta}}_k) - L_k(\boldsymbol{\theta}_k^*)$, $k = 1, \ldots, K$, otherwise the estimate of the correction $\mathfrak{c}(\alpha)$ can be too conservative. For instance, the Bonferroni correction would lead to the marginal confidence level $1 - \alpha/K$, which may be very conservative if K is large and the statistics $L_k(\widetilde{\boldsymbol{\theta}}_k) - L_k(\boldsymbol{\theta}_k^*)$ are highly correlated.

In Section 3.1 we suggest a multiplier bootstrap procedure, which performs the steps 1 and 2 described above. Theoretical justification of the procedure is given in Section

3.2. The proofs are based on several approximation bounds: non-asymptotic square-root Wilks theorem, simultaneous Gaussian approximation for ℓ_2-norms, Gaussian comparison, and simultaneous Gaussian anti-concentration inequality.

In the Sections 1.1, 1.2 above we introduced the 1-st subproblem: it is considered only one parametric model ($K = 1$) there. Chapter 2 studies a multiplier procedure in detail for this case. Chapter 3 extends the construction to simultaneous confidence estimation for a collection of parametric models, however, the results about simultaneous confidence sets do not follow directly from the 1-model case, and require the use different technical tools.

Below we illustrate the definitions (1.12)-(1.18) of the simultaneous likelihood-based confidence sets with two popular examples.

Example 1 (Simultaneous confidence band for local constant regression): Let $Y_1, \ldots, Y_n$ be independent random scalar observations and $X_1, \ldots, X_n$ some deterministic design points. Consider the following quadratic likelihood function reweighted with the kernel functions $K(\cdot)$:

$$L(\boldsymbol{\theta}, x, h) \stackrel{\text{def}}{=} -\frac{1}{2}\sum\nolimits_{i=1}^{n}(Y_i - \boldsymbol{\theta})^2 w_i(x,h),$$
$$w_i(x,h) \stackrel{\text{def}}{=} K(\{x - X_i\}/h),$$
$$K(x) \in [0,1], \ \int_{I\!R} K(x)dx = 1, \ K(x) = K(-x).$$

Here $h > 0$ denotes bandwidth, the local smoothing parameter. The target point and the local MLE read as:

$$\boldsymbol{\theta}^*(x,h) \stackrel{\text{def}}{=} \frac{\sum_{i=1}^n w_i(x,h) I\!E Y_i}{\sum_{i=1}^n w_i(x,h)}, \qquad \widetilde{\boldsymbol{\theta}}(x,h) \stackrel{\text{def}}{=} \frac{\sum_{i=1}^n w_i(x,h) Y_i}{\sum_{i=1}^n w_i(x,h)}.$$

$\widetilde{\boldsymbol{\theta}}(x,h)$ is also known as Nadaraya-Watson estimate. Fix a bandwidth h and consider the range of points $x_1, \ldots, x_K$. They yield K local constant models with the target parameters $\boldsymbol{\theta}_k^* \stackrel{\text{def}}{=} \boldsymbol{\theta}^*(x_k, h)$ and the likelihood functions $L_k(\boldsymbol{\theta}) \stackrel{\text{def}}{=} L(\boldsymbol{\theta}, x_k, h)$ for $k = 1, \ldots, K$. The confidence intervals for each model are defined as

$$\mathcal{E}_k(\mathfrak{z}, h) \stackrel{\text{def}}{=} \left\{ \boldsymbol{\theta} \in \Theta : L(\widetilde{\boldsymbol{\theta}}(x_k, h), x_k, h) - L(\boldsymbol{\theta}, x_k, h) \le \mathfrak{z}^2/2 \right\},$$

with the quintile functions $\mathfrak{z}_k(\alpha)$ and for the multiplicity correction $\mathfrak{c}(\alpha)$ from (1.15) and (1.18) they form the following simultaneous confidence band:

$$I\!P\left(\bigcap\nolimits_{k=1}^{K} \left\{ \boldsymbol{\theta}_k^* \in \mathcal{E}_k\big(\mathfrak{z}_k\left(\mathfrak{c}(\alpha)\right)\big) \right\} \right) \ge 1 - \alpha.$$

In Section 3.3 we provide results of numerical experiments for this model.

Example 2 (Multiple quantiles regression): Quantile regression is an important method of statistical analysis, widely used in various applications. It aims at estimating conditional quantile functions of a response variable, see Koenker (2005). Multiple quantiles regression model considers simultaneously several quantile regression functions based on a range of quantile indices, see e.g. Liu and Wu (2011); Qu (2008); He (1997). Let $Y_1, \ldots, Y_n$ be independent random scalar observations and $X_1, \ldots, X_n \in I\!R^d$ some deterministic design points, as in Example 1. Consider the following quantile regression models for $k = 1, \ldots, K$:

$$Y_i = g_k(X_i) + \varepsilon_{k,i}, \quad i = 1, \ldots, n,$$

where $g_k(\boldsymbol{x}) : I\!R^d \mapsto I\!R$ are unknown functions, the random values $\varepsilon_{k,1}, \ldots, \varepsilon_{k,n}$ are independent for each fixed k, and

$$I\!P(\varepsilon_{k,i} < 0) = \tau_k \quad \text{for all } i = 1, \ldots, n.$$

The range of quantile indices $\tau_1, \ldots, \tau_K \in (0,1)$ is known and fixed. We are interested in simultaneous parametric confidence sets for the functions $g_1(\cdot), \ldots, g_K(\cdot)$. Let $f_k(\boldsymbol{x}, \boldsymbol{\theta}) : I\!R^d \times I\!R^{p_k} \mapsto I\!R$ be known regression functions. Using the quantile regression approach by Koenker and Bassett Jr (1978), this problem can be treated with the quasi maximum likelihood method and the following log-likelihood functions:

$$\begin{aligned} L_k(\boldsymbol{\theta}) &= -\sum\nolimits_{i=1}^{n} \rho_{\tau_k}\left(Y_i - f_k(X_i, \boldsymbol{\theta})\right), \\ \rho_{\tau_k}(x) &\stackrel{\text{def}}{=} x\left(\tau_k - 1\!\!1\left\{x < 0\right\}\right). \end{aligned}$$

for $k = 1, \ldots, K$. This quasi log-likelihood function corresponds to the Asymmetric Laplace distribution with the density function $\tau_k(1-\tau_k)\mathrm{e}^{-\rho_{\tau_k}(x-a)}$. If $\tau = 1/2$, then $\rho_{1/2}(x) = |x|/2$ and $L(\boldsymbol{\theta}) = -\sum_{i=1}^{n} |Y_i - f_k(X_i, \boldsymbol{\theta})|/2$, which corresponds to the median regression.

1.5 Notation

$\|\cdot\|$	is the Euclidean norm for vectors and spectral norm for matrices;
$\|\cdot\|_{\max}$	is the maximum of absolute values of elements of a vector or of a matrix;
$\|\cdot\|_1$	is the sum of absolute values of elements of a vector or of a matrix.

$$\begin{aligned} \mathbf{1}_K &\stackrel{\text{def}}{=} (1, \ldots, 1)^\top \in I\!R^K, \\ p_{\text{sum}} &\stackrel{\text{def}}{=} p_1 + \cdots + p_K, \quad p_{\max} \stackrel{\text{def}}{=} \max_{1 \le k \le K} p_k. \end{aligned}$$

$\mathtt{C}$ is a generic constant. The value $\mathtt{x} > 0$ describes our tolerance level: all the results will be valid on a random set of probability $(1 - C\mathrm{e}^{-\mathtt{x}})$ for an explicit constant C. Everywhere we show how the error bounds depend on $p, p_1, \ldots, p_K$ and n for the case of the i.i.d. observations $Y_1, \ldots, Y_n$ and $\mathtt{x} \le \mathtt{C} \log n$. More details on it are given in Section A.3.1. In Section A.3 we also consider generalised linear model and linear quantile regression, and show for them the dependence on $p, p_1, \ldots, p_K$ and n of all the values appearing in main results and their conditions.

1.6 Organization of the thesis

- Chapter 2 considers the case of one parametric model, it includes:
  - Section 2.1 with description of the multiplier bootstrap procedure,
  - Sections 2.2, 2.3 with theoretical results justifying the procedure and Section 2.5 with the required conditions, in Section 2.5.3 we check the **(SmB)** condition for some popular models,
  - Section 2.4 with the results of numerical experiments for simulated data. We check the performance of the bootstrap procedure for small sample sizes (50 and 100) in the following cases:
    i. linear regression model
       a) without any misspecification,
       b) with misspecified heteroscedastic noise,
       c) with misspecified both regression function and noise, and with growing modelling bias,
    ii. logistic regression with growing modelling bias;
- Chapter 3 studies the problem of simultaneous confidence estimation, it includes:
  - Section 3.1 with description of the simultaneous multiplier bootstrap procedure,
  - Sections 3.2.1 and 3.2.2 with an overview of the theoretical approach and the theoretical results showing the bootstrap validity. The imposed conditions are given in Section 3.4,
  - Section 3.3 describing the results of numerical experiments. We construct simultaneous confidence corridors for local constant and local quadratic regressions using both bootstrap and Monte Carlo procedures. The quality

of the bootstrap procedure is checked by computing the effective simultaneous coverage probabilities of the bootstrap confidence sets. We also compare the widths of the confidence bands and the values of multiplicity correction obtained with bootstrap and with Monte Carlo procedures. The experiments confirm that the simultaneous bootstrap confidence sets and the bootstrap multiplicity correction become conservative if the local parametric model is considerably misspecified;

- The Appendix consists of Chapters A-D:
  - Chapter A collects the statements about non-asymptotic square-root Wilks approximations for $\boldsymbol{Y}$ and bootstrap worlds. In Section A.3 these results are specified for some common models: i.i.d. observations, generalised linear model and linear median regression, we also show the dependence of the non-asymptotic bounds on sample size and parameter's dimension.
  - Chapter B presents some useful statements about approximations between distributions of ℓ_2-norms of sums of independent random vectors. Namely, Gaussian approximation, Gaussian comparison and anti-concentration inequality.
  - Chapter C provides similar results as in Chapter B for joint distributions of sets of ℓ_2-norms of sums of independent random vectors.
  - Chapter D contains proofs of the main results.

The results presented in Chapters 2 and 3 are based on the papers Spokoiny and Zhilova (2015) and Zhilova (2015) respectively.

Chapter 2

Bootstrap likelihood-based confidence sets

A multiplier bootstrap procedure for construction of likelihood-based confidence sets is considered for finite samples and a possible model misspecification. Theoretical results justify the bootstrap validity for a small or moderate sample size and allow to control the impact of the parameter dimension p: the bootstrap approximation works if p^3/n is small. The main result about bootstrap validity continues to apply even if the underlying parametric model is misspecified under the so-called small modelling bias condition. In the case when the true model deviates significantly from the considered parametric family, the bootstrap procedure is still applicable but it becomes a bit conservative: the size of the constructed confidence sets is increased by the modelling bias. We illustrate the results with numerical examples for misspecified linear and logistic regression models.

2.1 Multiplier bootstrap procedure

Let $\ell_i(\boldsymbol{\theta})$ denote the parametric log-density of the i-th observation:

$$\ell_i(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \log\left(\frac{d\mathbb{P}_{\boldsymbol{\theta}}}{d\mu_0}(Y_i)\right),$$

then $L(\boldsymbol{\theta}) = \sum_{i=1}^n \ell_i(\boldsymbol{\theta})$. Consider i.i.d. scalar random variables u_i independent of $\boldsymbol{Y}$ with $\mathbb{E}u_i = 1$, $\operatorname{Var} u_i = 1$, $\mathbb{E}\exp(u_i) < \infty$ for all $i = 1, \ldots, n$. Multiply the summands of the likelihood function $L(\boldsymbol{\theta})$ with the new random variables:

$$L^\circ(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \sum\nolimits_{i=1}^n \ell_i(\boldsymbol{\theta})u_i,$$

then it holds $I\!E^\circ L^\circ(\boldsymbol{\theta}) = L(\boldsymbol{\theta})$, where $I\!E^\circ$ stands for the conditional expectation given $\boldsymbol{Y}$. Therefore, the quasi MLE for the $\boldsymbol{Y}$-world is a target parameter for the bootstrap world:

$$\operatorname{argmax}_{\boldsymbol{\theta}\in\Theta} I\!E^\circ L^\circ(\boldsymbol{\theta}) = \operatorname{argmax}_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta}) = \widetilde{\boldsymbol{\theta}}.$$

The corresponding quasi MLE under the conditional measure $I\!P^\circ$ is defined as

$$\widetilde{\boldsymbol{\theta}}^\circ \stackrel{\text{def}}{=} \operatorname{argmax}_{\boldsymbol{\theta}\in\Theta} L^\circ(\boldsymbol{\theta}).$$

The likelihood ratio statistic in the bootstrap world is equal to $L^\circ(\widetilde{\boldsymbol{\theta}}^\circ) - L^\circ(\widetilde{\boldsymbol{\theta}})$ in which all the entries are known including the function $L^\circ(\boldsymbol{\theta})$ and the arguments $\widetilde{\boldsymbol{\theta}}^\circ$, $\widetilde{\boldsymbol{\theta}}$.

Let $1-\alpha \in (0,1)$ be a known desirable confidence level of the set $\mathcal{E}(\mathfrak{z})$:

$$I\!P\left(\boldsymbol{\theta}^* \in \mathcal{E}(\mathfrak{z})\right) \geq 1-\alpha. \tag{2.1}$$

Here the parameter $\mathfrak{z} \geq 0$ determines the size of the confidence set. Define $\mathfrak{z}(\alpha)$ as the minimal possible value of $\mathfrak{z}$ such that (2.1) is fulfilled:

$$\mathfrak{z}(\alpha) \stackrel{\text{def}}{=} \inf\left\{\mathfrak{z} \geq 0\colon I\!P\left(L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*) > \mathfrak{z}^2/2\right) \leq \alpha\right\}. \tag{2.2}$$

For evaluating this value we apply the multiplier bootstrap procedure which replaces the unknown data distribution with the artificial bootstrap distribution given the observed sample. The target value $\mathfrak{z}(\alpha)$ is approximated by the value $\mathfrak{z}^\circ(\alpha)$ defined as the upper α-quantile of $\left\{2L^\circ(\widetilde{\boldsymbol{\theta}}^\circ) - 2L^\circ(\widetilde{\boldsymbol{\theta}})\right\}^{1/2}$:

$$\mathfrak{z}^\circ(\alpha) \stackrel{\text{def}}{=} \inf\left\{\mathfrak{z} \geq 0\colon I\!P^\circ\left(L^\circ(\widetilde{\boldsymbol{\theta}}^\circ) - L^\circ(\widetilde{\boldsymbol{\theta}}) > \mathfrak{z}^2/2\right) \leq \alpha\right\}. \tag{2.3}$$

Note that the bootstrap probability $I\!P^\circ$ and log-likelihood excess $L^\circ(\widetilde{\boldsymbol{\theta}}^\circ) - L^\circ(\widetilde{\boldsymbol{\theta}})$ depends on the data $\boldsymbol{Y}$ and thus, $\mathfrak{z}^\circ(\alpha)$ is random as well. Theoretical results of the next section justify the proposed approach.

2.2 Main results

Now we state the main results for the general set-up. The approximating error terms and the conditions are specified in Section A.3 for popular examples including i.i.d. observations, generalised regression model and linear quantile regression. Our first result claims that the random quantity $I\!P^\circ\left(L^\circ(\widetilde{\boldsymbol{\theta}}^\circ) - L^\circ(\widetilde{\boldsymbol{\theta}}) > \mathfrak{z}^2/2\right)$ is close in probability to the value $I\!P\left(L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*) > \mathfrak{z}^2/2\right)$ for a wide range of $\mathfrak{z}$-values.

Theorem 2.1. *Let the conditions of Section 2.5 be fulfilled, then it holds for* $\mathfrak{z} \geq \max\{2, \sqrt{p}\} + \mathtt{C}(p+\mathtt{x})/\sqrt{n}$ *with probability* $\geq 1 - 12\mathrm{e}^{-\mathtt{x}}$ *:*

$$\left| I\!\!P\left(L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*) > \mathfrak{z}^2/2\right) - I\!\!P^\circ\left(L^\circ(\widetilde{\boldsymbol{\theta}}^\circ) - L^\circ(\widetilde{\boldsymbol{\theta}}) > \mathfrak{z}^2/2\right)\right| \leq \Delta_{\mathrm{full}}\,.$$

The error term $\Delta_{\mathrm{full}} \leq \mathtt{C}\{(p+\mathtt{x})^3/n\}^{1/8}$ *in the case of i.i.d. model; see Section A.3.1. Explicit definition of the error term* Δ_{full} *is given in Section D.1.1, see* (D.9) *and* (D.11) *therein.*

The term Δ_{full} can be viewed as a sum of the error terms corresponding to each step in the scheme (1.6). The largest error term equal to $\mathtt{C}\{(p+\mathtt{x})^3/n\}^{1/8}$ is induced by GAR. This error rate is not always optimal for GAR, e.g. in the case of $p = 1$ or for the i.i.d. observations (see Remark B.2). In Theorems 2.3 and 2.5 the rate is $\mathtt{C}\{(p+\mathtt{x})^3/n\}^{1/2}$.

The next result can be viewed as "bootstrap validity".

Theorem 2.2 (Validity of the bootstrap under a small modelling bias)**.** *Assume the conditions of Theorem 2.1. Then for* $\alpha \leq 1 - 8\mathrm{e}^{-\mathtt{x}}$ *, it holds*

$$\left| I\!\!P\left(L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*) > (\mathfrak{z}^\circ(\alpha))^2/2\right) - \alpha\right| \leq \Delta_{\mathfrak{z},\,\mathrm{full}}\,. \tag{2.4}$$

The error term $\Delta_{\mathfrak{z},\,\mathrm{full}} \leq \mathtt{C}\{(p+\mathtt{x})^3/n\}^{1/8}$ *in the case of i.i.d. model; see Section A.3.1. For a precise description see* (D.17) *and* (D.18)*.*

In view of definition (1.1) of the likelihood-based confidence set Theorem 2.1 implies the following

Corollary 2.1 (Coverage probability error)**.** *Under the conditions of Theorem 2.2 it holds:*

$$\left| I\!\!P\left\{\boldsymbol{\theta}^* \in \mathcal{E}\left(\mathfrak{z}^\circ(\alpha)\right)\right\} - (1-\alpha)\right| \leq \Delta_{\mathfrak{z},\,\mathrm{full}}.$$

Remark 2.1 (Critical dimension)**.** The error term Δ_{full} depends on the ratio p^3/n. The bootstrap validity can be only stated if this ratio is small. The obtained error bound seems to be mainly of theoretical interest, because the condition "$(p^3/n)^{1/8}$ is small" may require a huge sample. However, it provides some qualitative information about the bootstrap behavior as the parameter dimension grows. Our numerical results show that the accuracy of bootstrap approximation is very reasonable in a variety of examples with $p \ll n$.

In the following theorem we consider the case of the scalar parameter $p = 1$. The obtained error rate is $1/\sqrt{n}$, which is sharper than $1/n^{1/8}$. Instead of the GAR for the Euclidean norm from Section B we use here Berry-Esseen theorem (see also Remark B.2).

Theorem 2.3 (The case of $p = 1$, using Berry-Esseen theorem)**.** *Let the conditions of Section 2.5 be fulfilled.*

1. *For* $\mathfrak{z} \geq 1 + \mathtt{C}(1+\mathtt{x})/\sqrt{n}$, *it holds with probability* $\geq 1 - 12\mathrm{e}^{-\mathtt{x}}$

$$\left| I\!\!P\left(L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*) > \mathfrak{z}^2/2\right) - I\!\!P^\circ\left(L^\circ(\widetilde{\boldsymbol{\theta}}^\circ) - L^\circ(\widetilde{\boldsymbol{\theta}}) > \mathfrak{z}^2/2\right)\right| \leq \Delta_{\text{B.E., full}}; \quad (2.5)$$

2. *For* $\alpha \leq 1 - 8\mathrm{e}^{-\mathtt{x}}$

$$\left| I\!\!P\left(L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*) > (\mathfrak{z}^\circ(\alpha))^2/2\right) - \alpha \right| \leq \Delta_{\text{B.E.}\,\mathfrak{z}\text{, full}}\,. \quad (2.6)$$

The error terms $\Delta_{\text{B.E., full}}, \Delta_{\text{B.E.}\,\mathfrak{z}\text{, full}} \leq \mathtt{C}(1+\mathtt{x})/\sqrt{n}$ *in the case A.3.1. Explicit definitions of* $\Delta_{\text{B.E., full}}$ *is given in* (D.20) *and* (D.21) *in Section D.1.1.*

Remark 2.2 (Bootstrap validity and weak convergence)**.** The standard way of proving the bootstrap validity is based on weak convergence arguments; see e.g. Mammen (1992), van der Vaart and Wellner (1996), Janssen and Pauls (2003), Chatterjee and Bose (2005). If the statistic $L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*)$ weakly converges to a χ^2-type distribution, one can state an asymptotic version of the results of Theorems 2.1, 2.3. Our way is based on a kind of non-asymptotic Gaussian approximation and Gaussian comparison for random vectors and allows to get explicit error terms.

Remark 2.3 (Use of Edgeworth expansion)**.** The classical results on confidence sets for the mean of population states the accuracy of order $1/n$ based on the second order Edgeworth expansion, see Hall (1992). Unfortunately, if the considered parametric model can be misspecified, even the leading term is affected by the modelling bias, and the use of Edgeworth expansion cannot help in improving the bootstrap accuracy.

Remark 2.4 (Choice of the weights)**.** In our construction, similarly to Chatterjee and Bose (2005), we apply a general distribution of the bootstrap weights u_i under some moment conditions. One particularly can use Gaussian multipliers as suggested by Chernozhukov et al. (2013a). This leads to the exact Gaussian distribution of the vectors $\boldsymbol{\xi}^\circ$ and is helpful to avoid one step of Gaussian approximation for these vectors.

Remark 2.5 (Skipping the Gaussian approximation step)**.** The biggest error term $\mathtt{C}\{(p+\mathtt{x})^3/n\}^{1/8}$ in Theorem 2.1 is induced by the Gaussian approximation step. In some particular cases the Gaussian approximation step can be avoided leading to better error bounds. For example, if the marginal score vectors $\nabla_{\boldsymbol{\theta}}\ell_i(\boldsymbol{\theta}^*)$ are normally distributed, and the random bootstrap weights are normal as well, $u_i \sim \mathcal{N}(1,1)$, then the vectors $\boldsymbol{\xi}$ and $\boldsymbol{\xi}^\circ$ are automatically normal, and the GAR step can be skipped.

If the marginal score vectors $\nabla_{\boldsymbol{\theta}}\ell_i(\boldsymbol{\theta}^*)$ are i.i.d. and symmetrically distributed (s.t. $\nabla_{\boldsymbol{\theta}}\ell_i(\boldsymbol{\theta}^*) \sim -\nabla_{\boldsymbol{\theta}}\ell_i(\boldsymbol{\theta}^*)$), and the centered bootstrap weights follow the Rademacher distribution ($u_i \sim 2Bernoulli(0.5)$), then the recent results by Arlot et al. (2010a) can be applied to show that the conditional distribution of $\|\boldsymbol{\xi}^\circ(\boldsymbol{\theta}^*)\|$ given the data is close to the distribution of $\|\boldsymbol{\xi}\|$. However, such methods require some special structural conditions on the underlying measure $I\!P$ like symmetricity or Gaussianity of the errors and may fail if these conditions are violated. It remains a challenging question how a nice performance of a general bootstrap procedure even for small or moderate samples can be explained.

Now we discuss the impact of modelling bias, which comes from a possible misspecification of the parametric model. As explained by the approximating diagram (1.6), the distance between the distributions of the likelihood ratio statistics can be characterized via the distance between two multivariate normal distributions. To state the result let us recall the definition of the full Fisher information matrix $D_0^2 \stackrel{\text{def}}{=} -\nabla^2_{\boldsymbol{\theta}} I\!E L(\boldsymbol{\theta}^*)$. For the matrices H_0^2 and B_0^2, given in (1.9) and (1.10), it holds $H_0^2 > B_0^2 \geq 0$. If the parametric assumption (1.1) is true or in the case of an i.i.d. sample $\boldsymbol{Y}$, $B_0^2 = 0$. Under the condition **(SmB)** $\|H_0^{-1}B_0^2H_0^{-1}\|$ enters linearly in the error term Δ_{full} in Theorem 2.1.

The first statement in Theorem 2.4 below says that the effective coverage probability of the confidence set based on the multiplier bootstrap is *larger* than the nominal coverage probability up to the error term $\Delta_{\text{b, full}} \leq \mathtt{C}\{(p+\mathtt{x})^3/n\}^{1/8}$. The inequalities in the second part of Theorem 2.4 prove the *conservativeness of the bootstrap quantiles*: the quantity $\sqrt{\text{tr}\{D_0^{-1}H_0^2D_0^{-1}\}} - \sqrt{\text{tr}\{D_0^{-1}(H_0^2 - B_0^2)D_0^{-1}\}} \geq 0$ increases with the growing modelling bias.

Theorem 2.4 (Performance of the bootstrap for a large modelling bias)**.** *Under the conditions of Section 2.5 except for* **(SmB)** *it holds for* $\mathfrak{z} \geq \max\{2, \sqrt{p}\} + \mathtt{C}(p+\mathtt{x})/\sqrt{n}$ *with probability* $\geq 1 - 14\mathrm{e}^{-\mathtt{x}}$

1. $$I\!P\left(L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*) > \mathfrak{z}^2/2\right) \leq I\!P^\circ\left(L^\circ(\widetilde{\boldsymbol{\theta}}^\circ) - L^\circ(\widetilde{\boldsymbol{\theta}}) > \mathfrak{z}^2/2\right) + \Delta_{\text{b, full}}.$$

2. $$\mathfrak{z}^\circ(\alpha) \geq \mathfrak{z}(\alpha + \Delta_{\text{b, full}}) + \sqrt{\text{tr}\{D_0^{-1}H_0^2D_0^{-1}\}} - \sqrt{\text{tr}\{D_0^{-1}(H_0^2 - B_0^2)D_0^{-1}\}} - \Delta_{\mathtt{qf},1},$$

   $$\mathfrak{z}^\circ(\alpha) \leq \mathfrak{z}(\alpha - \Delta_{\text{b, full}}) + \sqrt{\text{tr}\{D_0^{-1}H_0^2D_0^{-1}\}} - \sqrt{\text{tr}\{D_0^{-1}(H_0^2 - B_0^2)D_0^{-1}\}} + \Delta_{\mathtt{qf},2}.$$

The term $\Delta_{\mathrm{b,\,full}} \leq \mathtt{C}\{(p+\mathtt{x})^3/n\}^{1/8}$ *is given in* (D.23) *in Section D.1.2. The positive values* $\Delta_{\mathtt{qf},1}, \Delta_{\mathtt{qf},2}$ *are given in* (D.28), (D.27) *in Section D.1.2, they are bounded from above with* $(\mathfrak{a}^2 + \mathfrak{a}_B^2)(\sqrt{8\mathtt{x}p} + 6\mathtt{x})$ *for the constants* $\mathfrak{a}^2 > 0$, $\mathfrak{a}_B^2 \geq 0$ *from conditions* $(\boldsymbol{\mathcal{I}})$, $(\boldsymbol{\mathcal{I}_B})$.

Remark 2.6. There exists some literature on robust (and heteroscedasticity robust) bootstrap procedures; see e.g. Mammen (1993), Aerts and Claeskens (2001), Kline and Santos (2012). However, up to our knowledge there are no robust bootstrap procedures for the likelihood ratio statistic, most of the results compare the distribution of the estimator obtained from estimating equations, or Wald / score test statistics with their bootstrap counterparts in the i.i.d. setup. In our context this would correspond to the noise misspecification in the log-likelihood function and it is addressed automatically by the multiplier bootstrap. Our notion of modelling bias includes the situation when the target value $\boldsymbol{\theta}^*$ from (1.3) only defines a projection (the best parametric fit) of the data distribution. In particularly, the quantities $I\!\!E \nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}^*)$ for different i do not necessarily vanish yielding a significant modelling bias. Similar notion of misspecification is used in the literature on Generalized Method of Moments; see e.g. Hall (2005). Chapter 5 therein considers the hypothesis testing problem with two kinds of misspecification: local and non-local, which would correspond to our small and large modelling bias cases.

An interesting message of Theorem 2.4 is that the multiplier bootstrap procedure ensures a prescribed coverage level for this target value $\boldsymbol{\theta}^*$ even without small modelling bias restriction, however, in this case the method is somehow conservative because the modelling bias is transferred into the additional variance in the bootstrap world. The numerical experiments in Section 2.4 agree with this result.

2.3 Smoothed version of a quantile function

This section explains how to improve the accuracy of bootstrap approximation using a smoothed quantile function. The $(1-\alpha)$-quantile of $\sqrt{2L(\widetilde{\boldsymbol{\theta}}) - 2L(\boldsymbol{\theta}^*)}$ is defined as

$$\begin{aligned} \mathfrak{z}(\alpha) &\stackrel{\text{def}}{=} \inf\left\{\mathfrak{z} \geq 0 \colon I\!P\left(L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*) > \mathfrak{z}^2/2\right) \leq \alpha\right\} \\ &= \inf\left\{\mathfrak{z} \geq 0 \colon I\!E\, 1\!\!1\left\{L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*) > \mathfrak{z}^2/2\right\} \leq \alpha\right\}. \end{aligned}$$

Introduce for $x \geq 0$ and $z, \Delta > 0$ the following function

$$g_\Delta(x, z) \stackrel{\text{def}}{=} g\left(\frac{1}{2\Delta z}\left(x^2 - z^2\right)\right), \tag{2.7}$$

where $g(x)$ is a three times differentiable non-negative function, and grows monotonously from 0 to 1, $g(x) = 0$ for $x \le 0$ and $g(x) = 1$ for $x \ge 1$, therefore:

$$\mathbb{1}\{x > 1\} \le g(x) \le \mathbb{1}\{x > 0\} \le g(x+1).$$

An example of such function is given in (B.11). In holds

$$\mathbb{1}\{x - z > \Delta\} \le g_\Delta(x, z) \le \mathbb{1}(x - z > 0) \le g_\Delta(x, z + \Delta).$$

This approximation is used in the proofs of Theorems 2.1, 2.2 and 2.4 in the part of Gaussian approximation of Euclidean norm of a sum of independent vectors (see Section C.1) yielding the error rate $(p^3/n)^{1/8}$ in the final bound (Theorems 2.1, 2.2 and B.1). The next result shows that the use of a smoothed quantile function helps to improve the accuracy of bootstrap approximation: it becomes $(p^3/n)^{1/2}$ instead of $(p^3/n)^{1/8}$. The reason is that we do not need to account for the error induced by a smooth approximation of the indicator function.

Theorem 2.5 (Validity of the bootstrap in the smoothed case under **(SmB)** condition)**.** *Let the conditions of Section 2.5 be fulfilled. It holds for* $\mathfrak{z} \ge \max\{2, \sqrt{p}\} + \mathtt{C}(p + \mathtt{x})/\sqrt{n}$ *and* $\Delta \in (0, 0.22]$ *with probability* $\ge 1 - 12\mathrm{e}^{-\mathtt{x}}$ *:*

$$\left| I\!\!E g_\Delta \left(\sqrt{2L(\widetilde{\boldsymbol{\theta}}) - 2L(\boldsymbol{\theta}^*)}, \mathfrak{z} \right) - I\!\!E^\circ g_\Delta \left(\sqrt{2L^\circ(\widetilde{\boldsymbol{\theta}}^\circ) - 2L^\circ(\widetilde{\boldsymbol{\theta}})}, \mathfrak{z} \right) \right| \le \Delta_{\mathrm{sm}},$$

where $\Delta_{\mathrm{sm}} \le \mathtt{C}\{(p + \mathtt{x})^3/n\}^{1/2}\Delta^{-3}$ *in the case A.3.1. An explicit definition of* Δ_{sm} *is given in* (D.32)*,* (D.33) *in Section D.1.3.*

The modified bootstrap quantile function reads as

$$\mathfrak{z}^\circ_\Delta(\alpha) \stackrel{\mathrm{def}}{=} \inf\left\{ \mathfrak{z} \ge 0 \colon I\!\!E^\circ g_\Delta \left(\sqrt{2L^\circ(\widetilde{\boldsymbol{\theta}}^\circ) - 2L^\circ(\widetilde{\boldsymbol{\theta}})}, \mathfrak{z} \right) \le \alpha \right\}. \tag{2.8}$$

2.4 Numerical results

This section illustrates the performance of the multiplier bootstrap for some artificial examples. We especially aim to address the issues of noise misspecification and of increasing modelling bias. It should be mentioned that the obtained results are nicely consistent with the theoretical statements.

In all the experiments we took 10^4 data samples for estimation of the empirical c.d.f. of $\sqrt{2L(\widetilde{\boldsymbol{\theta}}) - 2L(\boldsymbol{\theta}^*)}$, and 10^4 $\{u_1, \ldots, u_n\}$ samples for each of the 10^4 data samples for the estimation of the quantiles of $\sqrt{2L^\circ(\widetilde{\boldsymbol{\theta}}^\circ) - 2L^\circ(\widetilde{\boldsymbol{\theta}})}$.

2.4.1 Computational error

Here we check numerically, how well the multiplier procedure works in the case of the correct model. Here the modelling bias term $\|H_0^{-1}B_0^2H_0^{-1}\|$ from the **(SmB)** condition equals to zero by its definition. Let the data come from the following model: $Y_i = \Psi_i^\top \boldsymbol{\theta}_0 + \varepsilon_i$, for $i = 1, \ldots, n$, where $\varepsilon_i \sim \mathcal{N}(0,1)$, $\Psi_i \stackrel{\text{def}}{=} \left(1, X_i, X_i^2, \ldots, X_i^{p-1}\right)^\top$, the design points $X_1, \ldots, X_n$ are equidistant on $[0,1]$, and the parameter vector $\boldsymbol{\theta}_0 = (1, \ldots, 1)^\top \in I\!R^p$. The true likelihood function is $L(\boldsymbol{\theta}) = -\sum_{i=1}^n (Y_i - \Psi_i^\top \boldsymbol{\theta})^2/2$. In this experiment we consider three cases: the scalar parameter $p = 1$, and the multivariate parameter $p = 3,\ 10$.

Table 2.1 shows the effective coverage probabilities of the quantiles estimated using the multiplier bootstrap. The second line contains the range of the nominal confidence levels: $0.99, \ldots, 0.75$. The first left column shows the sample size n and the second column - the parameter's dimension p. The third left column describes the distribution of the bootstrap weights: $2Bernoulli(0.5)$, $\mathcal{N}(1,1)$ or $exp(1)$. Below its 2-nd line the table contains the frequencies of the event: "the real likelihood ratio $\leq$ the quantile of the bootstrap likelihood ratio".

Table 2.1: Coverage probabilities for the correct model

			Confidence levels					
n	p	$\mathcal{L}(u_i)$	**0.99**	**0.95**	**0.90**	**0.85**	**0.80**	**0.75**
50	1	$2Bernoulli(0.5)$	0.986	0.942	0.892	0.838	0.792	0.745
		$\mathcal{N}(1,1)$	0.988	0.945	0.895	0.847	0.803	0.751
		$exp(1)$	0.988	0.942	0.885	0.833	0.784	0.729
50	3	$2Bernoulli(0.5)$	0.984	0.938	0.885	0.838	0.788	0.736
		$\mathcal{N}(1,1)$	0.994	0.949	0.897	0.844	0.789	0.736
		$exp(1)$	0.984	0.917	0.835	0.776	0.707	0.650
50	10	$2Bernoulli(0.5)$	0.975	0.923	0.866	0.813	0.764	0.715
		$\mathcal{N}(1,1)$	0.996	0.950	0.877	0.780	0.721	0.644
		$exp(1)$	0.952	0.827	0.710	0.617	0.541	0.473

2.4.2 Linear regression with misspecified heteroscedastic errors

Here we show on a linear regression model that the quality of the confidence sets obtained by the multiplier bootstrap procedure is not significantly deteriorated by

misspecified heteroscedastic errors. Let the data be defined as $Y_i = \Psi_i^\top \boldsymbol{\theta}_0 + \sigma_i \varepsilon_i$, $i = 1, \ldots, n$. The i.i.d. random variables $\varepsilon_i \sim \mathit{Laplace}(0, 2^{-1/2})$ are s.t. $I\!\!E(\varepsilon_i) = 0$, $\mathrm{Var}(\varepsilon_i) = 1$. The coefficients σ_i are deterministic: $\sigma_i \stackrel{\mathrm{def}}{=} 0.5\,\{4 - i \pmod 4\}$. The regressors Ψ_i are the same as in the experiment 2.4.1. The quasi-likelihood function is also the same as in the previous section: $L(\boldsymbol{\theta}) = -\sum_{i=1}^n (Y_i - \Psi_i^\top \boldsymbol{\theta})^2/2$, and it is misspecified, since it corresponds to $\sigma_i \varepsilon_i \sim \mathcal{N}(0,1)$. The target point $\boldsymbol{\theta}^* = \boldsymbol{\theta}_0$, therefore, the modelling bias term $\|H_0^{-1} B_0^2 H_0^{-1}\|$ from the **(SmB)** condition equals to zero.

Here we also consider three different parameter's dimensions: $p = 1, 3, 10$ with $\boldsymbol{\theta}_0 = (1, \ldots, 1)^\top \in I\!R^p$. Table 2.2 describes the 2-nd experiment's results similarly to the Table 2.1.

Table 2.2: Coverage probabilities for case of misspecified heteroscedastic noise

			Confidence levels					
n	p	$\mathcal{L}(u_i)$	**0.99**	**0.95**	**0.90**	**0.85**	**0.80**	**0.75**
50	1	$2Bernoulli(0.5)$	0.988	0.947	0.896	0.849	0.799	0.752
		$\mathcal{N}(1,1)$	0.990	0.949	0.893	0.844	0.794	0.746
		$exp(1)$	0.989	0.941	0.881	0.825	0.770	0.714
50	3	$2Bernoulli(0.5)$	0.984	0.937	0.885	0.834	0.788	0.739
		$\mathcal{N}(1,1)$	0.996	0.955	0.897	0.839	0.780	0.722
		$exp(1)$	0.988	0.924	0.846	0.765	0.701	0.634
50	10	$2Bernoulli(0.5)$	0.976	0.927	0.870	0.815	0.765	0.715
		$\mathcal{N}(1,1)$	0.998	0.959	0.891	0.810	0.731	0.655
		$exp(1)$	0.967	0.850	0.726	0.630	0.552	0.479
100	10	$2Bernoulli(0.5)$	0.985	0.935	0.885	0.833	0.781	0.733
		$\mathcal{N}(1,1)$	0.998	0.970	0.917	0.857	0.786	0.723
		$exp(1)$	0.989	0.921	0.826	0.741	0.663	0.591

One can see from the Tables 2.1 and 2.2 that the bootstrap procedure does a good job even for small or moderate samples like 50 or 100 if the parameter dimension is not too large. The results are stable w.r.t. the noise misspecification.

The Rademacher and Gaussian weights demonstrate nearly the same nice performance while the procedure with exponential weights tends to underestimate the real quantiles. This effect becomes especially prominent when the parameter dimension grows to 10.

2.4.3 Biased constant regression with misspecified errors

In the third experiment we consider biased regression with misspecified i.i.d. errors:

$$Y_i = \beta \sin(X_i) + \varepsilon_i, \quad \varepsilon_i \sim \mathit{Laplace}(0, 2^{-1/2}), \text{ i.i.d},$$
$$X_i \text{ are equidistant in } [0, 2\pi].$$

Taking the likelihood function $L(\boldsymbol{\theta}) = -\sum_{i=1}^{n}(Y_i - \boldsymbol{\theta})^2/2$ yields $\boldsymbol{\theta}^* = 0$. Therefore, the larger is the deterministic amplitude $\beta > 0$, the bigger is bias of the mean constant regression. The **(SmB)** condition reads as

$$\begin{aligned}\|H_0^{-1} B_0^2 H_0^{-1}\| &= 1 - \frac{\sum_{i=1}^{n} \operatorname{Var} Y_i}{\beta^2 \sum_{i=1}^{n} \sin^2(X_i) + \sum_{i=1}^{n} \operatorname{Var} Y_i} \\ &= 1 - \frac{1}{\beta^2 (n-1)/2n + 1} \\ &\le 1/\sqrt{n}.\end{aligned}$$

Consider the sample size $n = 50$, and two cases: $\beta = 0.25$ with fulfilled **(SmB)** condition and $\beta = 1.25$ when **(SmB)** does not hold. Table 2.3 shows that for the large bias quantiles yielded by the multiplier bootstrap are conservative. This

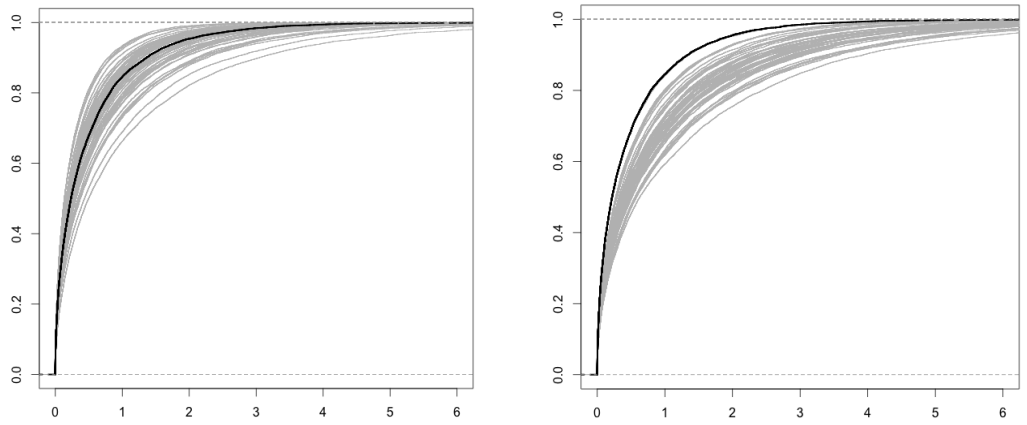
Table 2.3: Coverage probabilities for the noise-misspecified biased regression

			Confidence levels					
n	$\mathcal{L}(u_i)$	β	**0.99**	**0.95**	**0.90**	**0.85**	**0.80**	**0.75**
50	$\mathcal{N}(1,1)$	0.25	0.98	0.94	0.89	0.84	0.79	0.74
		1.25	1.0	0.99	0.97	0.94	0.91	0.87

conservative property of the multiplier bootstrap quantiles is also illustrated with the graphs in Figure 2.1. They show the empirical distribution functions of the likelihood ratio statistics $L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*)$ and $L^\circ(\widetilde{\boldsymbol{\theta}}^\circ) - L^\circ(\widetilde{\boldsymbol{\theta}})$ for $\beta = 0.25$ and $\beta = 1.25$. On the right graph for $\beta = 1.25$ the empirical distribution functions for the bootstrap case are smaller than the one for the $\boldsymbol{Y}$ case. It means that for the large bias the bootstrap quantiles are bigger than the $\boldsymbol{Y}$ quantiles, which increases the diameter of the confidence set based on the bootstrap quantiles. This confidence set remains valid, since it still contains the true parameter with a given confidence level.

Figure 2.2 shows the growth of the difference between the quantiles of $L^\circ(\widetilde{\boldsymbol{\theta}}^\circ) - L^\circ(\widetilde{\boldsymbol{\theta}})$ and $L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*)$ with increasing β for the range of the confidence levels: $0.75, 0.8, \ldots, 0.99$.

Figure 2.1: Empirical distribution functions of the likelihood ratios

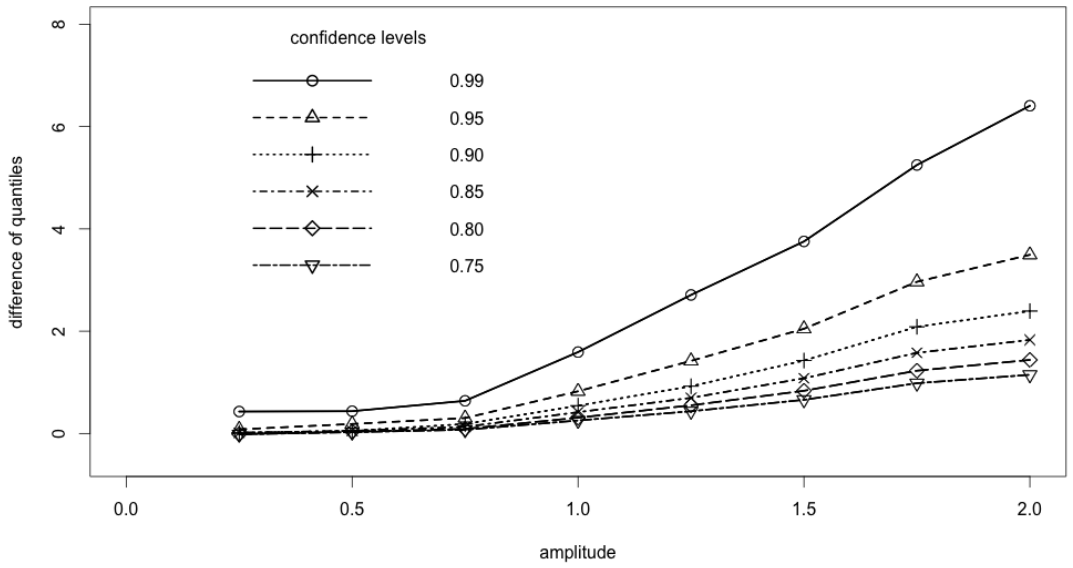


$Y_i = 0.25\sin(X_i) + Lap(0, 2^{-1/2})$, $n = 50$ $\quad$ $Y_i = 1.25\sin(X_i) + Lap(0, 2^{-1/2})$, $n = 50$

— empirical distribution function of $L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*)$ estimated with 10^4 $\boldsymbol{Y}$ samples

— 50 empirical distribution functions of $L^\bullet(\widetilde{\boldsymbol{\theta}}^\bullet) - L^\bullet(\widetilde{\boldsymbol{\theta}})$ estimated with 10^4 $\{u_i\} \sim exp(1)$ samples

Figure 2.2: The difference $\big($ "Bootstrap quantile" – "$\boldsymbol{Y}$-quantile" $\big)$ growing with modelling bias



2.4.4 Logistic regression with bias

In this example we consider logistic regression. Let the data come from the following distribution:

$$Y_i \sim Bernoulli(\beta X_i), \quad X_i \text{ are equidistant in } [0,2], \quad \beta \in (0, 1/2].$$

Consider the likelihood function corresponding to the i.i.d. observations:

$$L(\boldsymbol{\theta}) = \sum\nolimits_{i=1}^{n} \left\{ Y_i \boldsymbol{\theta} - \log(1 + \mathrm{e}^{\boldsymbol{\theta}}) \right\}.$$

By definition (1.3) $\boldsymbol{\theta}^* = \log\{\beta/(1-\beta)\}$, bigger values of β induce larger modelling bias. Indeed, the **(SmB)** condition reads as:

$$\|H_0^{-1} B_0^2 H_0^{-1}\| = \frac{\beta^2 \sum_{i=1}^n (X_i - 1)^2}{n\beta^2 + \beta(1-2\beta)\sum_{i=1}^n X_i} \tag{2.9}$$

$$= \frac{\beta}{1-\beta} \cdot \frac{n+1}{3(n-1)} \tag{2.10}$$

$$\leq 1/\sqrt{n}. \tag{2.11}$$

The graphs on Figure 2.3 demonstrate the conservativeness of bootstrap quantiles. Here we consider two cases: $\beta = 0.1$ and $\beta = 0.5$. Similarly to the Example 2.4.3 in the case of the bigger β on the right graph of Figure 2.3 the empirical distribution functions of $L^\circ(\widetilde{\boldsymbol{\theta}}^\circ) - L^\circ(\widetilde{\boldsymbol{\theta}})$ are smaller than the one for $L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*)$.

2.5 Conditions

Here we state the conditions required for the main results. The conditions in Section 3.4.1 come from the general finite sample theory by Spokoiny (2012a), they are required for the results of Sections A.1 and A.2. The conditions in Section 3.4.2 are required to prove the results on multiplier bootstrap from Section 2.1. In Section A.3 we verify these conditions in detail for several examples: i.i.d. observations, generalised linear model and linear quantile regression.

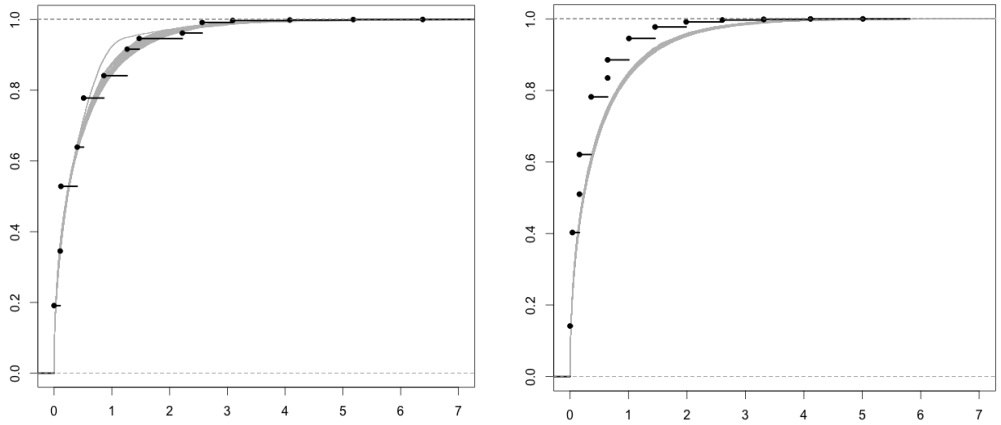
2.5.1 Basic conditions

Introduce the stochastic part of the likelihood process: $\zeta(\boldsymbol{\theta}) \stackrel{\text{def}}{=} L(\boldsymbol{\theta}) - I\!\!E L(\boldsymbol{\theta})$, and its marginal summand: $\zeta_i(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \ell_i(\boldsymbol{\theta}) - I\!\!E \ell_i(\boldsymbol{\theta})$.

(ED_0) *There exist a positive-definite symmetric matrix* V_0^2 *and constants* $\mathtt{g} > 0, \nu_0 \geq 1$ *such that* $\operatorname{Var}\{\nabla_{\boldsymbol{\theta}} \zeta(\boldsymbol{\theta}^*)\} \leq V_0^2$ *and*

$$\sup_{\boldsymbol{\gamma} \in I\!R^p} \log I\!\!E \exp\left\{ \lambda \frac{\boldsymbol{\gamma}^\top \nabla_{\boldsymbol{\theta}} \zeta(\boldsymbol{\theta}^*)}{\|V_0 \boldsymbol{\gamma}\|} \right\} \leq \nu_0^2 \lambda^2 / 2, \qquad |\lambda| \leq \mathtt{g}.$$

Figure 2.3: Empirical distribution functions of the likelihood ratios for logistic regression



$Y_i \sim Bernoulli(0.1X_i), n = 50$ $Y_i \sim Bernoulli(0.5X_i), n = 50$

— empirical distribution function of $L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*)$ estimated with 10^4 $\boldsymbol{Y}$ samples

— 50 empirical distribution functions of $L^{\circ}(\widetilde{\boldsymbol{\theta}}^{\circ}) - L^{\circ}(\widetilde{\boldsymbol{\theta}})$ estimated with 10^4 $\{u_i\} \sim exp(1)$ samples

($\boldsymbol{ED_2}$) *There exist a constant* $\omega \geq 0$ *and for each* $\mathtt{r} > 0$ *a constant* $\mathtt{g}_2(\mathtt{r})$ *such that it holds for all* $\boldsymbol{\theta} \in \Theta_0(\mathtt{r})$ *and for* $j = 1, 2$

$$\sup_{\substack{\boldsymbol{\gamma}_j \in I\!R^p \\ \|\boldsymbol{\gamma}_j\| \leq 1}} \log I\!E \exp\left\{ \frac{\lambda}{\omega} \boldsymbol{\gamma}_1^{\top} D_0^{-1} \nabla^2_{\boldsymbol{\theta}} \zeta(\boldsymbol{\theta}) D_0^{-1} \boldsymbol{\gamma}_2 \right\} \leq \nu_0^2 \lambda^2 / 2, \qquad |\lambda| \leq \mathtt{g}_2(\mathtt{r}).$$

($\boldsymbol{\mathcal{L}_0}$) *For each* $\mathtt{r} \in [0, \mathtt{r}_0]$ *($\mathtt{r}_0$ comes from condition* (A.2) *of Theorem A.1) there exists a constant* $\delta(\mathtt{r}) \in [0, 1/2]$ *s.t. for all* $\boldsymbol{\theta} \in \Theta_0(\mathtt{r})$ *it holds*

$$\|D_0^{-1} D^2(\boldsymbol{\theta}) D_0^{-1} - \boldsymbol{I}_p\| \leq \delta(\mathtt{r}),$$

where $D^2(\boldsymbol{\theta}) \stackrel{\text{def}}{=} -\nabla^2_{\boldsymbol{\theta}} I\!E L(\boldsymbol{\theta})$, $\Theta_0(\mathtt{r}) \stackrel{\text{def}}{=} \{\boldsymbol{\theta} : \|D_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \leq \mathtt{r}\}$.

($\boldsymbol{\mathcal{I}}$) *There exists a constant* $\mathtt{a} > 0$ *s.t.* $\mathtt{a}^2 D_0^2 \geq V_0^2$.

($\boldsymbol{\mathcal{L}\mathtt{r}}$) *For each* $\mathtt{r} > \mathtt{r}_0$ *there exists a value* $\mathtt{b}(\mathtt{r}) > 0$ *s.t.* $\mathtt{rb}(\mathtt{r}) \to +\infty$ *for* $\mathtt{r} \to +\infty$ *and* $\forall \boldsymbol{\theta} : \|D_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| = \mathtt{r}$ *it holds*

$$-2\{I\!E L(\boldsymbol{\theta}) - I\!E L(\boldsymbol{\theta}^*)\} \geq \mathtt{r}^2 \mathtt{b}(\mathtt{r}).$$

2.5.2 Conditions required for the bootstrap validity

(SmB) *There exists a constant* $\delta^2_{\mathrm{smb}} \in [0, 1/8]$ *such that it holds for the matrices* H_0^2, B_0^2 *defined in* (1.9) *and* (1.10).

$$\|H_0^{-1} B_0^2 H_0^{-1}\| \leq \delta^2_{\mathrm{smb}} \leq \mathtt{C} p n^{-1/2}.$$

($\boldsymbol{ED_{2m}}$) *For each* $\mathtt{r} > 0$, $i = 1, \ldots, n$, $j = 1, 2$ *and for all* $\boldsymbol{\theta} \in \Theta_0(\mathtt{r})$ *it holds for the values* $\omega \geq 0$ *and* $\mathtt{g}_2(\mathtt{r})$ *from the condition* **($\boldsymbol{ED_2}$)**:

$$\sup_{\substack{\boldsymbol{\gamma}_j \in I\!R^p \\ \|\boldsymbol{\gamma}_j\| \leq 1}} \log I\!E \exp\left\{ \frac{\lambda}{\omega} \boldsymbol{\gamma}_1^\top D_0^{-1} \nabla^2_{\boldsymbol{\theta}} \zeta_i(\boldsymbol{\theta}) D_0^{-1} \boldsymbol{\gamma}_2 \right\} \leq \frac{\nu_0^2 \lambda^2}{2n}, \qquad |\lambda| \leq \mathtt{g}_2(\mathtt{r}).$$

($\boldsymbol{\mathcal{L}_{0m}}$) *For each* $\mathtt{r} > 0$, $i = 1, \ldots, n$ *and for all* $\boldsymbol{\theta} \in \Theta_0(\mathtt{r})$ *there exists a constant* $\mathtt{C}_m(\mathtt{r}) \geq 0$ *such that*

$$\|D_0^{-1} \nabla^2_{\boldsymbol{\theta}} I\!E \ell_i(\boldsymbol{\theta}) D_0^{-1}\| \leq \mathtt{C}_m(\mathtt{r}) n^{-1}.$$

($\boldsymbol{\mathcal{I}_B}$) *There exists a constant* $\mathfrak{a}_B^2 \geq 0$ *s.t.* $\mathfrak{a}_B^2 D_0^2 \geq B_0^2$.

($\boldsymbol{SD_1}$) *There exists a constant* $0 \leq \delta_v^2 \leq \mathtt{C} p/n$. *such that it holds for all* $i = 1, \ldots, n$ *with exponentially high probability*

$$\left\| H_0^{-1} \left\{ \nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}^*) \nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}^*)^\top - I\!E \left[\nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}^*) \nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}^*)^\top \right] \right\} H_0^{-1} \right\| \leq \delta_v^2.$$

($\boldsymbol{Eb}$) *The bootstrap weights* u_i *are i.i.d., independent of the data* $\boldsymbol{Y}$, *and*

$$I\!E u_i = 1, \quad \operatorname{Var} u_i = 1,$$
$$\log I\!E \exp\{\lambda(u_i - 1)\} \leq \nu_0^2 \lambda^2 / 2, \qquad |\lambda| \leq \mathtt{g}.$$

2.5.3 Small modelling bias condition for some models

Here we consider what becomes with the condition **(SmB)** for some particular models. If the observations $Y_1, \ldots, Y_n$ are i.i.d., then $\nabla_{\boldsymbol{\theta}} I\!E L(\boldsymbol{\theta}^*) = n \nabla_{\boldsymbol{\theta}} I\!E \ell_i(\boldsymbol{\theta}^*) = 0$, and $B_0^2 = 0$. The next example is the generalized linear model: the parametric probability distribution family $\{I\!P_\upsilon\}$ is an exponential family with a canonical parametrisation. The log-density for this family can be expressed as

$$\ell(\upsilon) = y\upsilon - h(\upsilon)$$

for a convex function $h(\cdot)$. Table 2.4 provides some examples of $\{I\!P_\upsilon\}$ and $h(\cdot)$. Tak-

Table 2.4: Examples of the GLM

$I\!P_\upsilon$	$h(\upsilon)$	$h'(\upsilon)$ (natural parameter)
$\mathcal{N}(\upsilon, 1)$	$\upsilon^2/2$	υ
$Exp(-\upsilon)$	$-\log(-\upsilon)$	$-1/\upsilon$
$Pois(\mathrm{e}^\upsilon)$	e^υ	e^υ
$Binom\left(1, \frac{\mathrm{e}^\upsilon}{\mathrm{e}^\upsilon+1}\right)$	$\log(\mathrm{e}^\upsilon + 1)$	$\frac{\mathrm{e}^\upsilon}{\mathrm{e}^\upsilon+1}$

ing $\{I\!P_\upsilon\}$ as a parametric family and $\Psi_i^\top \boldsymbol{\theta}$ as linear predictors for some deterministic regressors $\Psi_i \in I\!R^p$ yields the following quasi log-likehood function:

$$L(\boldsymbol{\theta}) = \sum\nolimits_{i=1}^n \left\{ Y_i \Psi_i^\top \boldsymbol{\theta} - h(\Psi_i^\top \boldsymbol{\theta}) \right\}.$$

It holds

$$\|H_0^{-1} B_0^2 H_0^{-1}\| \le 1 - \min_{1 \le i \le n} \frac{\operatorname{Var} Y_i}{\operatorname{Var} Y_i + \left\{ I\!E Y_i - h'(\Psi_i^\top \boldsymbol{\theta}^*) \right\}^2} \in [0, 1).$$

It is important that $I\!E_{\boldsymbol{\theta}^*} Y_i = h'(\Psi_i^\top \boldsymbol{\theta}^*)$, i.e. in the case of the correct parametric model $I\!P \in \{I\!P_\upsilon\}$ the modelling bias is indeed equal to zero.

Now let us consider the linear qauntile regression. Let the observations $Y_1, \ldots, Y_n$ be scalar, and the design points $X_1, \ldots, X_n$ be deterministic. Let $\tau \in (0, 1)$ denote a fixed known quantile level. The object of estimation is a quantile function $q_\tau(x)$ s.t.

$$I\!P\big(Y_i < q_\tau(X_i)\big) = \tau \quad \forall\, i = 1, \ldots, n.$$

Using the quantile regression approach by Koenker and Bassett Jr (1978), this problem can be treated with quasi maximum likelihood method and the following log-likelihood function:

$$\begin{aligned} L(\boldsymbol{\theta}) &= -\sum\nolimits_{i=1}^n \rho_\tau \left(Y_i - \Psi_i^\top \boldsymbol{\theta} \right), \\ \rho_\tau(x) &\stackrel{\text{def}}{=} x \left(\tau - 1\!\!1 \{x < 0\} \right), \end{aligned} \tag{2.12}$$

where $\Psi_i \in I\!R^p$ are known regressors. This log-likelihood function corresponds to asymmetric Laplace distribution with the density $\tau(1-\tau)\mathrm{e}^{-\rho_\tau(x-a)}$. It holds

$$\begin{aligned} &\|H_0^{-1} B_0^2 H_0^{-1}\| \\ &\le 1 - \min_{1 \le i \le n} \frac{\operatorname{Var}\left(\tau - 1\!\!1\big\{Y_i - \Psi_i^\top \boldsymbol{\theta}^* < 0\big\}\right)}{\operatorname{Var}\left(\tau - 1\!\!1\big\{Y_i - \Psi_i^\top \boldsymbol{\theta}^* < 0\big\}\right) + \left(\tau - I\!P\big\{Y_i - \Psi_i^\top \boldsymbol{\theta}^* < 0\big\}\right)^2}. \end{aligned}$$

If $I\!P\big\{Y_i - \Psi_i^\top \boldsymbol{\theta}^* < 0\big\} \equiv \tau$, then the right side of the last inequality is equal to zero.

Chapter 3

Simultaneous bootstrap confidence sets

This chapter studies a problem of construction of simultaneous likelihood-based confidence sets. We consider a simultaneous multiplier bootstrap procedure for estimating the quantiles of the joint distribution of the likelihood ratio statistics, and for adjusting the confidence level for multiplicity. Theoretical results state the bootstrap validity in the following setting: the sample size n is fixed, the maximal parameter dimension $p_{\max}$ and the number of considered parametric models K are s.t. $(\log K)^{12} p_{\max}^3 / n$ is small. We also consider the situation when the parametric models are misspecified. If the models' misspecification is significant, then the bootstrap critical values exceed the true ones and the simultaneous bootstrap confidence set becomes conservative. Numerical experiments for local constant and local quadratic regressions illustrate the theoretical results.

3.1 Simultaneous multiplier bootstrap procedure

Let $\ell_{i,k}(\boldsymbol{\theta})$ denote the log-density from the k-th parametric distribution family evaluated at the i-th observation:

$$\ell_{i,k}(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \log \left(\frac{d\mathbb{P}_k(\boldsymbol{\theta})}{d\mu_0}(Y_i) \right), \tag{3.1}$$

then due to independence of $Y_1, \ldots, Y_n$

$$L_k(\boldsymbol{\theta}) = \sum\nolimits_{i=1}^{n} \ell_{i,k}(\boldsymbol{\theta}) \quad \forall\, k = 1, \ldots, K.$$

Consider i.i.d. scalar random variables u_i independent of the data $\boldsymbol{Y}$, s.t. $I\!\!E u_i = 1$, $\operatorname{Var} u_i = 1$, $I\!\!E \exp(u_i) < \infty$ (e.g. $u_i \sim \mathcal{N}(1,1)$ or $u_i \sim exp(1)$ or $u_i \sim 2Bernoulli(0.5)$). Multiply the summands of the likelihood function $L_k(\boldsymbol{\theta})$ with the new random variables:

$$L_k^\circ(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \sum\nolimits_{i=1}^n \ell_{i,k}(\boldsymbol{\theta}) u_i, \tag{3.2}$$

then it holds $I\!\!E^\circ L_k^\circ(\boldsymbol{\theta}) = L_k(\boldsymbol{\theta})$, where $I\!\!E^\circ$ stands for the conditional expectation given $\boldsymbol{Y}$.

Therefore, the quasi MLE for the $\boldsymbol{Y}$-world is a target parameter for the bootstrap world for each $k = 1, \ldots, K$:

$$\operatorname{argmax}_{\boldsymbol{\theta} \in \Theta_k} I\!\!E^\circ L_k^\circ(\boldsymbol{\theta}) = \operatorname{argmax}_{\boldsymbol{\theta} \in \Theta_k} L_k(\boldsymbol{\theta}) = \widetilde{\boldsymbol{\theta}}_k.$$

The corresponding bootstrap MLE is:

$$\widetilde{\boldsymbol{\theta}}_k^\circ \stackrel{\text{def}}{=} \operatorname{argmax}_{\boldsymbol{\theta} \in \Theta_k} L_k^\circ(\boldsymbol{\theta}).$$

The k-th likelihood ratio statistic in the bootstrap world equals to $L_k^\circ(\widetilde{\boldsymbol{\theta}}_k^\circ) - L_k^\circ(\widetilde{\boldsymbol{\theta}}_k)$, where all the elements: the function $L_k^\circ(\boldsymbol{\theta})$ and the arguments $\widetilde{\boldsymbol{\theta}}_k^\circ$, $\widetilde{\boldsymbol{\theta}}_k$ are known and available for computation. This means, that given the data $\boldsymbol{Y}$, one can estimate the distribution or quantiles of the statistic $L_k^\circ(\widetilde{\boldsymbol{\theta}}_k^\circ) - L_k^\circ(\widetilde{\boldsymbol{\theta}}_k)$ by generating many independent samples of the bootstrap weights $u_1, \ldots, u_n$ and computing with them the bootstrap likelihood ratio.

Let us introduce similarly to (1.15) the $(1-\alpha)$-quantile for the bootstrap square-root likelihood ratio statistic:

$$\mathfrak{z}_k^\circ(\alpha) \stackrel{\text{def}}{=} \inf\left\{ \mathfrak{z} \geq 0 : I\!\!P^\circ\left(L_k^\circ(\widetilde{\boldsymbol{\theta}}_k^\circ) - L_k^\circ(\widetilde{\boldsymbol{\theta}}_k) > \mathfrak{z}^2/2 \right) \leq \alpha \right\}, \tag{3.3}$$

here $I\!\!P^\circ$ denotes probability measure conditional on the data $\boldsymbol{Y}$, therefore, $\mathfrak{z}_k^\circ(\alpha)$ is a random value dependent on $\boldsymbol{Y}$.

In Chapter 2 we consider the case of a single parametric model ($K = 1$), and showed that the bootstrap quantile $\mathfrak{z}_k^\circ(\alpha)$ is close to the true one $\mathfrak{z}_k(\alpha)$ under the **(SmB)** condition, which is fulfilled when the true distribution is close to the parametric family or when the observations are i.i.d. When the **(SmB)** condition does not hold, the bootstrap quantile is still valid, however, it becomes conservative. Therefore, for each fixed $k = 1, \ldots, K$ the bootstrap quantiles $\mathfrak{z}_k^\circ(\alpha)$ are rather good estimates for the true unknown ones $\mathfrak{z}_k(\alpha)$, however, they are still "pointwise" in k, i.e. the confidence bounds (1.16) hold for each k separately. Our goal here is to

estimate $\mathfrak{z}_1(\alpha), \ldots, \mathfrak{z}_K(\alpha)$ and $\mathfrak{c}(\alpha)$ according to (1.17) and (1.18). Let us introduce the bootstrap correction for multiplicity:

$$\mathfrak{c}^{\circ}(\alpha) \stackrel{\text{def}}{=} \sup\left\{c \in (0,\alpha] : I\!P^{\circ}\left(\bigcup\nolimits_{k=1}^{K}\left\{\sqrt{2L_k^{\circ}(\widetilde{\boldsymbol{\theta}}_k^{\circ}) - 2L_k^{\circ}(\widetilde{\boldsymbol{\theta}}_k)} > \mathfrak{z}_k^{\circ}(c)\right\}\right) \le \alpha\right\}. \quad (3.4)$$

By its definition $\mathfrak{c}^{\circ}(\alpha)$ depends on the random sample $\boldsymbol{Y}$.

The multiplier bootstrap procedure below explains how to estimate the bootstrap quantile functions $\mathfrak{z}_k^{\circ}(\mathfrak{c}^{\circ}(\alpha))$ corrected for multiplicity.

The simultaneous bootstrap procedure:

Input: The data $\boldsymbol{Y}$ (as in (1.11)) and a fixed confidence level $(1-\alpha) \in (0,1)$.

Step 1: Generate B independent samples of i.i.d. bootstrap weights $\{u_1^{(b)}, \ldots, u_n^{(b)}\}$, $b = 1, \ldots, B$. For the bootstrap likelihood processes

$$L_k^{\circ(b)}(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \sum\nolimits_{i=1}^{n} \ell_{i,k}(\boldsymbol{\theta}) u_i^{(b)} \quad (3.5)$$

compute the bootstrap likelihood ratios $L_k^{\circ(b)}(\boldsymbol{\theta}_k^{\circ(b)}) - L_k^{\circ(b)}(\widetilde{\boldsymbol{\theta}}_k)$. For each fixed b the bootstrap likelihoods $L_1^{\circ(b)}(\boldsymbol{\theta}), \ldots, L_K^{\circ(b)}(\boldsymbol{\theta})$ are computed using the same bootstrap sample $\{u_i^{(b)}\}$, s.t. the i-th summand $\ell_{i,k}(\boldsymbol{\theta})$ is always multiplied with the i-th weight $u_i^{(b)}$ as in (3.5).

Step 2: Estimate the marginal quantile functions $\mathfrak{z}_k^{\circ}(\alpha)$ defined in (3.3) separately for each $k = 1, \ldots, K$, using B bootstrap realisations of $L_k^{\circ}(\widetilde{\boldsymbol{\theta}}_k^{\circ}) - L_k^{\circ}(\widetilde{\boldsymbol{\theta}}_k)$ from **Step 1**.

Step 3: Find by an iterative procedure the maximum value $c \in (0,\alpha]$ s.t.

$$I\!P^{\circ}\left(\bigcup\nolimits_{k=1}^{K}\left\{\sqrt{2L_k^{\circ}(\widetilde{\boldsymbol{\theta}}_k^{\circ}) - 2L_k^{\circ}(\widetilde{\boldsymbol{\theta}}_k)} \ge \mathfrak{z}_k^{\circ}(c)\right\}\right) \le \alpha.$$

Otput: The resulting critical values are $\mathfrak{z}_k^{\circ}(c)$, $k = 1, \ldots, K$.

Remark 3.1. The requirement in Step 1 to use the same bootstrap sample $\{u_i^{(b)}\}$ for generation of the bootstrap likelihood ratios $L_k^{\circ(b)}(\boldsymbol{\theta}_k^{\circ(b)}) - L_k^{\circ(b)}(\widetilde{\boldsymbol{\theta}}_k)$, $k = 1, \ldots, K$ allows to preserve the correlation structure between the ratios and, therefore, to make a sharper simultaneous adjustment in Step 3.

This procedure is justified theoretically in the next section.

3.2 Theoretical justification of the bootstrap procedure

Before stating the main results in Section 3.2.2 we introduce in Section 3.2.1 the basic ingredients of the proofs. The general scheme of the theoretical approach here is taken from the case of one parametric model (Section 1.3, scheme (1.6)). Here we extend that approach for the case of simultaneous confidence estimation for a collection of parametric models.

3.2.1 Overview of the theoretical approach

For justification of the described multiplier bootstrap procedure for simultaneous inference it has to be checked that the joint distributions of the sets of likelihood ratio statistics $\left\{L_k(\widetilde{\boldsymbol{\theta}}_k) - L_k(\boldsymbol{\theta}_k^*) : k = 1, \ldots, K\right\}$ and $\left\{L_k^\circ(\widetilde{\boldsymbol{\theta}}_k^\circ) - L_k^\circ(\widetilde{\boldsymbol{\theta}}_k) : k = 1, \ldots, K\right\}$ are close to each other. These joint distributions are approximated using several non-asymptotic steps given in the following scheme:

$$
\begin{array}{lccccc}
 & & \overset{\text{uniform sq-Wilks theorem}}{} & & \overset{\text{joint Gauss. approx. \& anti-concentr.}^*}{} & \\
\boldsymbol{Y}\text{-world:} & \sqrt{2L_k(\widetilde{\boldsymbol{\theta}}_k) - 2L_k(\boldsymbol{\theta}_k^*)} & \underset{\frac{p_k+\log K}{\sqrt{n}}}{\approx} & \|\boldsymbol{\xi}_k\| & \approx & \|\overline{\boldsymbol{\xi}}_k\| \\
 & & & \bigcap_{1\le k\le K} & & \wr\wr\ w \ \overset{\text{simultaneous Gauss. compar.}^\dagger}{} \\
\text{Bootstrap world:} & \sqrt{2L_k^\circ(\widetilde{\boldsymbol{\theta}}_k^\circ) - 2L_k^\circ(\widetilde{\boldsymbol{\theta}}_k)} & \underset{\frac{p_k+\log K}{\sqrt{n}}}{\approx} & \|\overline{\boldsymbol{\xi}}_k^\circ\| & \approx & \|\overline{\boldsymbol{\xi}}_k^\circ\|
\end{array}
\tag{3.6}
$$

* the accuracy of these approximating steps is $\mathtt{C}\left\{\frac{p_{\max}^3}{n}\log^9(K)\log^3(np_{\text{sum}})\right\}^{1/8}$;

† Gaussian comparison step yields an approximation error proportional to $\widehat{\delta}_{\text{smb}}^2\left(\frac{p_{\max}^3}{n}\right)^{1/4} p_{\max}\log^2(K)\log^{3/4}(np_{\text{sum}})$, where $\widehat{\delta}_{\text{smb}}^2$ comes from condition $\widehat{\textbf{(SmB)}}$, see also (3.9) below.

Here $\boldsymbol{\xi}_k$ and $\boldsymbol{\xi}_k^\circ$ denote normalized score vectors for the $\boldsymbol{Y}$ and bootstrap likelihood processes:

$$
\boldsymbol{\xi}_k \overset{\text{def}}{=} D_k^{-1}\nabla_{\boldsymbol{\theta}} L_k(\boldsymbol{\theta}_k^*), \qquad \boldsymbol{\xi}_k^\circ \overset{\text{def}}{=} \boldsymbol{\xi}_k^\circ(\boldsymbol{\theta}_k^*) \overset{\text{def}}{=} D_k^{-1}\nabla_{\boldsymbol{\theta}} L_k^\circ(\boldsymbol{\theta}_k^*), \tag{3.7}
$$

D_k^2 is the full Fisher information matrix for the corresponding k-th likelihood:

$$
D_k^2 \overset{\text{def}}{=} -\nabla_{\boldsymbol{\theta}}^2 I\!\!E L_k(\boldsymbol{\theta}_k^*).
$$

$\overline{\boldsymbol{\xi}}_k \sim \mathcal{N}(0, \operatorname{Var} \boldsymbol{\xi}_k)$ and $\overline{\boldsymbol{\xi}}^{\circ}_k \sim \mathcal{N}(0, \operatorname{Var}^{\circ} \boldsymbol{\xi}^{\circ}_k)$ denote approximating Gaussian vectors, which have the same covariance matrices as $\boldsymbol{\xi}_k$ and $\boldsymbol{\xi}^{\circ}_k$. Moreover the vectors $\left(\overline{\boldsymbol{\xi}}_1^{\top}, \ldots, \overline{\boldsymbol{\xi}}_K^{\top}\right)^{\top}$ and $\left(\overline{\boldsymbol{\xi}}^{\circ\top}_1, \ldots, \overline{\boldsymbol{\xi}}^{\circ\top}_K\right)^{\top}$ are normally distributed and have the same covariance matrices as the vectors $\left(\boldsymbol{\xi}_1^{\top}, \ldots, \boldsymbol{\xi}_K^{\top}\right)^{\top}$ and $\left(\boldsymbol{\xi}^{\circ\top}_1, \ldots, \boldsymbol{\xi}^{\circ\top}_K\right)^{\top}$ correspondingly. $\operatorname{Var}^{\circ}$ and $\operatorname{Cov}^{\circ}$ denote variance and covariance operators w.r.t. the probability measure $I\!P^{\circ}$ conditional on $\boldsymbol{Y}$.

The first two approximating steps: square root Wilks and Gaussian approximations are performed in parallel for both $\boldsymbol{Y}$ and bootstrap worlds, which is shown in the corresponding lines of the scheme (3.6). The two worlds are connected in the last step: Gaussian comparison for ℓ_2-norms of Gaussian vectors. All the approximations are performed simultaneously for K parametric models.

Let us consider each step in more details. Non-asymptotic square-root Wilks approximation result had been obtained recently by Spokoiny (2012a, 2013). It says that for a fixed sample size and misspecified parametric assumption: $I\!P \notin \{I\!P_k\}$, it holds with exponentially high probablity:

$$\left| \sqrt{2\{L_k(\widetilde{\boldsymbol{\theta}}_k) - L_k(\boldsymbol{\theta}^*_k)\}} - \|\boldsymbol{\xi}_k\| \right| \leq \Delta_{k,\mathrm{W}} \simeq \frac{p_k}{\sqrt{n}},$$

here the index k is fixed, i.e. this statement is for one parametric model. The precise statement of this result is given in Section A.1, and its simultaneous version – in Section A.4. The approximating value $\|\boldsymbol{\xi}_k\|$ is ℓ_2-norm of the score vector $\boldsymbol{\xi}_k$ given in (3.7). The next approximating step is between the joint distributions of $\|\boldsymbol{\xi}_1\|, \ldots, \|\boldsymbol{\xi}_K\|$ and $\|\overline{\boldsymbol{\xi}}_1\|, \ldots, \|\overline{\boldsymbol{\xi}}_K\|$. This is done in Section C.1 for general centered random vectors under bounded exponential moments assumptions. The main tools for the simultaneous Gaussian approximation are: Lindeberg's telescopic sum, smooth maximum function and three times differentiable approximation of the indicator function $\mathbb{1}\{x \in I\!R : x > 0\}$. The simultaneous anti-concentration inequality for the ℓ_2-norms of Gaussian vectors is obtained in Section C.3. The result is based on approximation of the ℓ_2-norm with a maximum over a finite grid on a hypersphere, and on the anti-concentration inequality for maxima of a Gaussian random vector by Chernozhukov et al. (2014c). The same approximating steps are performed for the bootstrap world, the square-root bootstrap Wilks approximation is given in Sections A.2, A.4. The last step in the scheme (3.6) is comparison of the joint distributions of the sets of ℓ_2-norms of Gaussian vectors: $\|\overline{\boldsymbol{\xi}}_1\|, \ldots, \|\overline{\boldsymbol{\xi}}_K\|$ and $\|\overline{\boldsymbol{\xi}}^{\circ}_1\|, \ldots, \|\overline{\boldsymbol{\xi}}^{\circ}_K\|$ by Slepian interpolation (see Section C.2 for the result in a general setting). The error of

approximation is proportional to

$$\max_{1\le k_1,k_2\le K} \left\|\mathrm{Cov}(\boldsymbol{\xi}_{k_1},\boldsymbol{\xi}_{k_2}) - \mathrm{Cov}^{\circ}(\boldsymbol{\xi}^{\circ}_{k_1},\boldsymbol{\xi}^{\circ}_{k_2})\right\|_{\max}. \tag{3.8}$$

It is shown, using Bernstein matrix inequality (Sections D.2.1 and D.2.2), that the value (3.8) is bounded from above (up to a constant) on a random set of dominating probability with

$$\max_{1\le k\le K} \left\|H_k^{-1}B_k^2H_k^{-1}\right\| \le \widehat{\delta}^2_{\mathrm{smb}} \tag{3.9}$$

for

$$\begin{aligned} B_k^2 &\stackrel{\mathrm{def}}{=} \sum\nolimits_{i=1}^n I\!\!E\left\{\nabla_{\boldsymbol{\theta}}\ell_{i,k}(\boldsymbol{\theta}_k^*)\right\} I\!\!E\left\{\nabla_{\boldsymbol{\theta}}\ell_{i,k}(\boldsymbol{\theta}_k^*)\right\}^{\top}, \\ H_k^2 &\stackrel{\mathrm{def}}{=} \sum\nolimits_{i=1}^n I\!\!E\left\{\nabla_{\boldsymbol{\theta}}\ell_{i,k}(\boldsymbol{\theta}_k^*)\nabla_{\boldsymbol{\theta}}\ell_{i,k}(\boldsymbol{\theta}_k^*)^{\top}\right\}. \end{aligned} \tag{3.10}$$

The value $\left\|H_k^{-1}B_k^2H_k^{-1}\right\|$ is responsible for the modelling bias of the k-th model. If the parametric family $\{I\!P_k(\boldsymbol{\theta})\}$ contains the true distribution $I\!P$ or if the observations $\boldsymbol{Y}_i$ are i.i.d., then B_k^2 equals to zero. Condition $\widehat{\textbf{(SmB)}}$ assumes that all the values $\left\|H_k^{-1}B_k^2H_k^{-1}\right\|$ are rather small.

3.2.2 Main results

The following theorem shows the closeness of the joint cumulative distribution functions (c.d.f-s.) of $\left\{\sqrt{2L_k(\widetilde{\boldsymbol{\theta}}_k) - 2L_k(\boldsymbol{\theta}_k^*)}, k=1,\ldots,K\right\}$ and $\left\{\sqrt{2L_k^{\circ}(\widetilde{\boldsymbol{\theta}}_k^{\circ}) - 2L_k^{\circ}(\widetilde{\boldsymbol{\theta}}_k)}, k=1,\ldots,K\right\}$. The approximating error term Δ_{total} equals to a sum of the errors from all the steps in the scheme (3.6).

Theorem 3.1. *Under the conditions of Section 3.4 it holds with probability* $\ge 1-12\mathrm{e}^{-\mathtt{x}}$ *for* $z_k \ge C\sqrt{p_k}$ *and a constant* $1 \le C < 2$

$$\begin{aligned} \Bigg| I\!P\left(\bigcup\nolimits_{k=1}^K \left\{\sqrt{2L_k(\widetilde{\boldsymbol{\theta}}_k) - 2L_k(\boldsymbol{\theta}_k^*)} > z_k\right\}\right) \\ - I\!P^{\circ}\left(\bigcup\nolimits_{k=1}^K \left\{\sqrt{2L_k^{\circ}(\widetilde{\boldsymbol{\theta}}_k^{\circ}) - 2L_k^{\circ}(\widetilde{\boldsymbol{\theta}}_k)} > z_k\right\}\right)\Bigg| \le \Delta_{\mathrm{total}}. \end{aligned}$$

The approximating total error $\Delta_{\mathrm{total}} \ge 0$ *is deterministic and in the case of i.i.d. observations (see Section A.3.1) it holds:*

$$\Delta_{\mathrm{total}} \le \mathtt{C}\left(\frac{p_{\max}^3}{n}\right)^{1/8} \log^{9/8}(K)\log^{3/8}(np_{\mathrm{sum}})\left\{\left(\widehat{\mathfrak{a}}^2+\widehat{\mathfrak{a}}_B^2\right)\left(1+\delta^2_{\widehat{\nu}}(\mathtt{x})\right)\right\}^{3/8}, \tag{3.11}$$

where the deterministic terms $\widehat{\mathfrak{a}}^2, \widehat{\mathfrak{a}}_B^2$ *and* $\delta^2_{\widehat{\nu}}(\mathtt{x})$ *come from the conditions* $\boldsymbol{(\mathcal{I}_k)}$, $\boldsymbol{(\mathcal{I}_{B,k})}$ *and* $\widehat{\boldsymbol{(SD_1)}}$. Δ_{total} *is defined in* (D.40).

Remark 3.2. The obtained approximation bound is mainly of theoretical interest, although it shows the impact of $p_{\max}$, K and n on the quality of the bootstrap procedure. For more details on the error term see Remark C.1.

The next theorem justifies the bootstrap procedure under the **(S$\widehat{\text{mB}}$)** condition. The theorem says that the bootstrap quantile functions $\mathfrak{z}_k^\circ(\cdot)$ with the bootstrap-corrected for multiplicity confidence levels $1 - \mathfrak{c}^\circ(\alpha)$ can be used for construction of the simultaneous confidence set in the $\boldsymbol{Y}$-world.

Theorem 3.2 (Bootstrap validity for a small modelling bias)**.** *Assume the conditions of Theorem 3.1, and* $\mathfrak{c}(\alpha), 0.5\mathfrak{c}^\circ(\alpha) \geq \Delta_{\text{full},\max}$*, then for* $\alpha \leq 1 - 8\mathrm{e}^{-\mathtt{x}}$ *it holds with probability* $1 - 12\mathrm{e}^{-\mathtt{x}}$

$$\mathbb{P}\left(\bigcup\nolimits_{k=1}^{K}\left\{\sqrt{2L_k(\widetilde{\boldsymbol{\theta}}_k) - 2L_k(\boldsymbol{\theta}_k^*)} \geq \mathfrak{z}_k^\circ\left(\mathfrak{c}^\circ(\alpha) - 2\Delta_{\text{full},\max}\right)\right\}\right) - \alpha \leq \Delta_{\mathfrak{z},\text{total}},$$
$$\mathbb{P}\left(\bigcup\nolimits_{k=1}^{K}\left\{\sqrt{2L_k(\widetilde{\boldsymbol{\theta}}_k) - 2L_k(\boldsymbol{\theta}_k^*)} \geq \mathfrak{z}_k^\circ\left(\mathfrak{c}^\circ(\alpha) + 2\Delta_{\text{full},\max}\right)\right\}\right) - \alpha \geq -\Delta_{\mathfrak{z},\text{total}},$$

where $\Delta_{\text{full},\max} \leq \mathtt{C}\{(p_{\max} + \mathtt{x})^3/n\}^{1/8}$ *in the case of i.i.d. observations (see Section A.3.1), and* $\Delta_{\mathfrak{z},\text{total}} \leq 3\Delta_{\text{total}}$*; their explicit definitions are given in* (D.46) *and* (D.49)*. Moreover*

$$\mathfrak{c}^\circ(\alpha) \leq \mathfrak{c}\left(\alpha + \Delta_{\mathfrak{c}}\right) + \Delta_{\text{full},\max},$$
$$\mathfrak{c}^\circ(\alpha) \geq \mathfrak{c}\left(\alpha - \Delta_{\mathfrak{c}}\right) - \Delta_{\text{full},\max},$$

for $0 \leq \Delta_{\mathfrak{c}} \leq 2\Delta_{\text{total}}$*, defined in* (D.50)*.*

The following theorem does not assume the **(S$\widehat{\text{mB}}$)** condition to be fulfilled. It turns out that in this case the bootstrap procedure becomes conservative, and the bootstrap critical values corrected for the multiplicity $\mathfrak{z}_k^\circ\left(\mathfrak{c}^\circ(\alpha)\right)$ are increased with the modelling bias $\sqrt{\operatorname{tr}\{D_k^{-1}H_k^2D_k^{-1}\}} - \sqrt{\operatorname{tr}\{D_k^{-1}(H_k^2 - B_k^2)D_k^{-1}\}}$, therefore, the confidence set based on the bootstrap estimates can be conservative.

Theorem 3.3 (Bootstrap conservativeness for a large modelling bias)**.** *Under the conditions of Section 3.4 except for* **(S$\widehat{\text{mB}}$)** *it holds with probability* $\geq 1 - 14\mathrm{e}^{-\mathtt{x}}$ *for* $z_k \geq C\sqrt{p_k}$ *and a constant* $1 \leq C < 2$

$$\mathbb{P}\left(\bigcup\nolimits_{k=1}^{K}\left\{\sqrt{2L_k(\widetilde{\boldsymbol{\theta}}_k) - 2L_k(\boldsymbol{\theta}_k^*)} > z_k\right\}\right)$$
$$\leq \mathbb{P}^\circ\left(\bigcup\nolimits_{k=1}^{K}\left\{\sqrt{2L_k^\circ(\widetilde{\boldsymbol{\theta}}_k^\circ) - 2L_k^\circ(\widetilde{\boldsymbol{\theta}}_k)} > z_k\right\}\right) + \Delta_{\text{b},\text{total}}.$$

The deterministic value $\Delta_{\mathrm{b,\,total}} \in [0, \Delta_{\mathrm{total}}]$ *(see* (3.11) *in the case A.3.1). Moreover, the bootstrap-corrected for multiplicity confidence level* $1 - \mathfrak{c}^{\circ}(\alpha)$ *is conservative in comparison with the true corrected confidence level:*

$$1 - \mathfrak{c}^{\circ}(\alpha) \geq 1 - \mathfrak{c}\,(\alpha + \Delta_{\mathrm{b},\mathfrak{c}}) - \Delta_{\mathrm{full,\,max}},$$

and it holds for all $k = 1, \ldots, K$ *and* $\alpha \leq 1 - 8\mathrm{e}^{-\mathtt{x}}$

$$\begin{aligned}\mathfrak{z}_k^{\circ}\left(\mathfrak{c}^{\circ}(\alpha)\right) \geq\ & \mathfrak{z}_k\left(\mathfrak{c}\,(\alpha + \Delta_{\mathrm{b},\mathfrak{c}}) + \Delta_{\mathrm{full,\,max}}\right) \\ & + \sqrt{\mathrm{tr}\{D_k^{-1} H_k^2 D_k^{-1}\}} - \sqrt{\mathrm{tr}\{D_k^{-1}(H_k^2 - B_k^2) D_k^{-1}\}} - \Delta_{\mathtt{qf},1,k},\end{aligned}$$

for $0 \leq \Delta_{\mathrm{b},\mathfrak{c}} \leq 2\Delta_{\mathrm{total}}$*, defined in* (D.53)*, and the positive value* $\Delta_{\mathtt{qf},1,k}$ *is bounded from above with* $(\mathfrak{a}_k^2 + \mathfrak{a}_{B,k}^2)(\sqrt{8\mathtt{x}p_k} + 6\mathtt{x})$ *for the constants* $\mathfrak{a}_k^2 > 0$, $\mathfrak{a}_{B,k}^2 \geq 0$ *from conditions* $(\boldsymbol{\mathcal{I}_k})$, $(\boldsymbol{\mathcal{I}_{B,k}})$.

The $\widehat{\textbf{(SmB)}}$ condition is automatically fulfilled if all the parametric models are correct or in the case of i.i.d. observations. This condition is checked for generalised linear model and linear quantile regression in Section 2.5.3.

3.3 Numerical experiments

Here we check the performance of the bootstrap procedure by constructing simultaneous confidence sets based on the local constant and local quadratic estimates, the former one is also known as Nadaraya-Watson estimate Nadaraya (1964); Watson (1964). Let $Y_1, \ldots, Y_n$ be independent random scalar observations and $X_1, \ldots, X_n$ some deterministic design points. In Sections 3.3.1-3.3.3 below we introduce the models and the data, Sections 3.3.4-3.3.6 present the results of the experiments.

3.3.1 Local constant regression

Consider the following quadratic likelihood function reweighted with the kernel functions $K(\cdot)$:

$$\begin{gathered}L(\boldsymbol{\theta}, x, h) \stackrel{\mathrm{def}}{=} -\frac{1}{2}\sum\nolimits_{i=1}^{n}(Y_i - \boldsymbol{\theta})^2 w_i(x,h), \\ w_i(x,h) \stackrel{\mathrm{def}}{=} K(\{x - X_i\}/h), \\ K(x) \in [0,1],\ \int_{I\!R} K(x)dx = 1,\ K(x) = K(-x).\end{gathered}$$

Here $h > 0$ denotes bandwidth, the local smoothing parameter. The target point and the local MLE read as:

$$\boldsymbol{\theta}^*(x,h) \stackrel{\mathrm{def}}{=} \frac{\sum_{i=1}^n w_i(x,h) I\!E Y_i}{\sum_{i=1}^n w_i(x,h)}, \qquad \widetilde{\boldsymbol{\theta}}(x,h) \stackrel{\mathrm{def}}{=} \frac{\sum_{i=1}^n w_i(x,h) Y_i}{\sum_{i=1}^n w_i(x,h)}.$$

Let us fix a bandwidth h and consider the range of points $x_1, \ldots, x_K$. They yield K local constant models with the target parameters $\boldsymbol{\theta}_k^* \stackrel{\text{def}}{=} \boldsymbol{\theta}^*(x_k, h)$ and the likelihood functions $L_k(\boldsymbol{\theta}) \stackrel{\text{def}}{=} L(\boldsymbol{\theta}, x_k, h)$ for $k = 1, \ldots, K$.

The bootstrap local likelihood function is defined similarly to the global one (3.2), by reweighting $L(\boldsymbol{\theta}, x, h)$ with the bootstrap multipliers $u_1, \ldots, u_n$:

$$L_k^\circ(\boldsymbol{\theta}) \stackrel{\text{def}}{=} L^\circ(\boldsymbol{\theta}, x_k, h) \stackrel{\text{def}}{=} -\frac{1}{2}\sum\nolimits_{i=1}^n (Y_i - \boldsymbol{\theta})^2 w_i(x_k, h) u_i,$$
$$\widetilde{\boldsymbol{\theta}}_k^\circ \stackrel{\text{def}}{=} \widetilde{\boldsymbol{\theta}}^\circ(x_k, h) \stackrel{\text{def}}{=} \frac{\sum_{i=1}^n w_i(x_k, h) u_i Y_i}{\sum_{i=1}^n w_i(x_k, h) u_i}.$$

3.3.2 Local quadratic regression

Here the local likelihood function reads as

$$L(\boldsymbol{\theta}, x, h) \stackrel{\text{def}}{=} -\frac{1}{2}\sum\nolimits_{i=1}^n (Y_i - \Psi_i^\top \boldsymbol{\theta})^2 w_i(x, h),$$
$$\boldsymbol{\theta}, \Psi_i \in I\!R^3, \quad \Psi_i \stackrel{\text{def}}{=} \left(1, X_i, X_i^2\right)^\top,$$

and

$$\boldsymbol{\theta}^*(x, h) \stackrel{\text{def}}{=} \left(\Psi W(x, h) \Psi^\top\right)^{-1} \Psi W(x, h) I\!E \boldsymbol{Y},$$
$$\widetilde{\boldsymbol{\theta}}(x, h) \stackrel{\text{def}}{=} \left(\Psi W(x, h) \Psi^\top\right)^{-1} \Psi W(x, h) \boldsymbol{Y},$$

where

$$\boldsymbol{Y} \stackrel{\text{def}}{=} (Y_1, \ldots, Y_n)^\top, \quad \Psi \stackrel{\text{def}}{=} (\Psi_1, \ldots, \Psi_n) \in I\!R^{3 \times n},$$
$$W(x, h) \stackrel{\text{def}}{=} \operatorname{diag}\{w_1(x, h), \ldots, w_n(x, h)\}.$$

And similarly for the bootstrap objects

$$L^\circ(\boldsymbol{\theta}, x, h) \stackrel{\text{def}}{=} -\frac{1}{2}\sum\nolimits_{i=1}^n (Y_i - \Psi_i^\top \boldsymbol{\theta})^2 w_i(x, h) u_i,$$
$$\widetilde{\boldsymbol{\theta}}^\circ(x, h) \stackrel{\text{def}}{=} \left(\Psi U W(x, h) \Psi^\top\right)^{-1} \Psi U W(x, h) \boldsymbol{Y},$$

for $U \stackrel{\text{def}}{=} \operatorname{diag}\{u_1, \ldots, u_n\}$.

3.3.3 Simulated data

In the numerical experiments we constructed two 90% simultaneous confidence bands: using Monte Carlo (MC) samples and bootstrap procedure with Gaussian weights ($u_i \sim \mathcal{N}(1, 1)$), in each case we used 10^4 $\{Y_i\}$ and 10^4 $\{u_i\}$ independent samples.

The sample size $n = 400$. $K(x)$ is Epanechnikov's kernel function. The independent random observations Y_i are generated as follows:

$$Y_i = f(X_i) + \mathcal{N}(0,1), \quad X_i \text{ are equidistant on } [0,1], \tag{3.12}$$

$$f(x) = \begin{cases} 5, & x \in [0, 0.25] \cup [0.65, 1]; \\ 5 + 3.8\{1 - 100(x - 0.35)^2\}, & x \in [0.25, 0.45]; \\ 5 - 3.8\{1 - 100(x - 0.55)^2\}, & x \in [0.45, 0.65]. \end{cases} \tag{3.13}$$

The number of local models $K = 71$, the points $x_1, \ldots, x_{71}$ are equidistant on $[0,1]$. For the bandwidth we considered two cases: $h = 0.12$ and $h = 0.3$.

3.3.4 Effect of the modelling bias on a width of a bootstrap confidence band

The function $f(x)$ defined in (3.13) should yield a considerable modelling bias for both mean constant and mean quadratic estimators. Figures 3.1, 3.2 demonstrate that the bootstrap confidence bands become conservative (i.e. wider than the MC confidence band) when the local model is misspecified. The top graphs on Figures 3.1, 3.2 show the 90% confidence bands, the middle graphs show their width, and the bottom graphs show the value of the modelling bias for $K = 71$ local models (see formulas (3.14) and (3.15) below). For the local constant estimate (Figure 3.1) the width of the bootstrap confidence sets is considerably increased by the modelling bias when $x \in [0.25, 0.65]$. In this case case the expression for the modelling bias term for the k-th model (see also **(SmB)** condition) reads as:

$$\begin{aligned} \left|H_k^{-1} B_k^2 H_k^{-1}\right| &= \frac{\sum_{i=1}^n \left\{I\!\!E Y_i - \boldsymbol{\theta}^*(x_k)\right\}^2 w_i^2(x_k, h)}{\sum_{i=1}^n I\!\!E \left\{Y_i - \boldsymbol{\theta}^*(x_k)\right\}^2 w_i^2(x_k, h)} \\ &= 1 - \left(1 + \frac{\sum_{i=1}^n w_i^2(x_k, h) \left\{f(X_i) - \boldsymbol{\theta}^*(x_k)\right\}^2}{\sum_{i=1}^n w_i^2(x_k, h)}\right)^{-1}. \end{aligned} \tag{3.14}$$

And for the local quadratic estimate it holds:

$$\left\|H_k^{-1} B_k^2 H_k^{-1}\right\| = \left\|\boldsymbol{I}_p - H_k^{-1} \left\{\sum\nolimits_{i=1}^n \Psi_i \Psi_i^\top w_i^2(x_k, h)\right\} H_k^{-1}\right\|, \tag{3.15}$$

where $\boldsymbol{I}_p$ is the identity matrix of dimension $p \times p$ (here $p = 3$), and

$$\begin{aligned} H_k^2 &= \sum\nolimits_{i=1}^n \Psi_i \Psi_i^\top w_i^2(x_k, h) I\!\!E \left\{Y_i - \boldsymbol{\theta}^*(x_k)\right\}^2 \\ &= \sum\nolimits_{i=1}^n \Psi_i \Psi_i^\top w_i^2(x_k, h) \left\{f(X_i) - \boldsymbol{\theta}^*(x_k)\right\}^2 + \sum\nolimits_{i=1}^n \Psi_i \Psi_i^\top w_i^2(x_k, h). \end{aligned} \tag{3.16}$$

Therefore, if $\max_{1 \le k \le K} \left\{f(X_i) - \boldsymbol{\theta}^*(x_k)\right\}^2 = 0$, then $\left\|H_k^{-1} B_k^2 H_k^{-1}\right\| = 0$. On the Figure 3.1 both the modelling bias and the difference between the widths of the

bootstrap and MC confidence bands are close to zero in the regions where the true function $f(x)$ is constant. On Figure 3.2 the modelling bias for $h = 0.12$ is overall smaller than the corresponding value on Figure 3.1. For the bigger bandwidth $h = 0.3$ the modelling biases on Figures 3.1 and 3.2 are comparable with each other.

Thus the numerical experiment is consistent with the theoretical results from Section 3.2.2, and confirm that in the case when a (local) parametric model is close to the true distribution the simultaneous bootstrap confidence set is valid. Otherwise the bootstrap procedure is conservative: the modelling bias widens the simultaneous bootstrap confidence set.

3.3.5 Effective coverage probability (local constant estimate)

In this part of the experiment we check the bootstrap validity by computing the effective coverage probability values. This requires to perform many independent experiments: for each of independent 5000 $\{Y_i\} \sim$ (3.12) samples we took 10^4 independent bootstrap samples $\{u_i\} \sim \mathcal{N}(1,1)$, and constructed simultaneous bootstrap confidence sets for a range of confidence levels. The second row of Table 3.1 contains this range $(1-\alpha) = 0.95, 0.9, \ldots, 0.5$. The third and the fourth rows of Table 3.1 show the frequencies of the event

$$\max_{1\le k\le K} \left\{ L_k(\widetilde{\boldsymbol{\theta}}_k) - L_k(\boldsymbol{\theta}_k^*) - \mathfrak{z}_k^\circ(\mathfrak{c}^\circ(\alpha)) \right\} \le 0$$

among 5000 data samples, for the bandwidths $h = 0.12, 0.3$, and for the range of $(1-\alpha)$. The results show that the bootstrap procedure is rather conservative for both $h = 0.12$ and $h = 0.3$, however, the larger bandwidth yields bigger coverage probabilities.

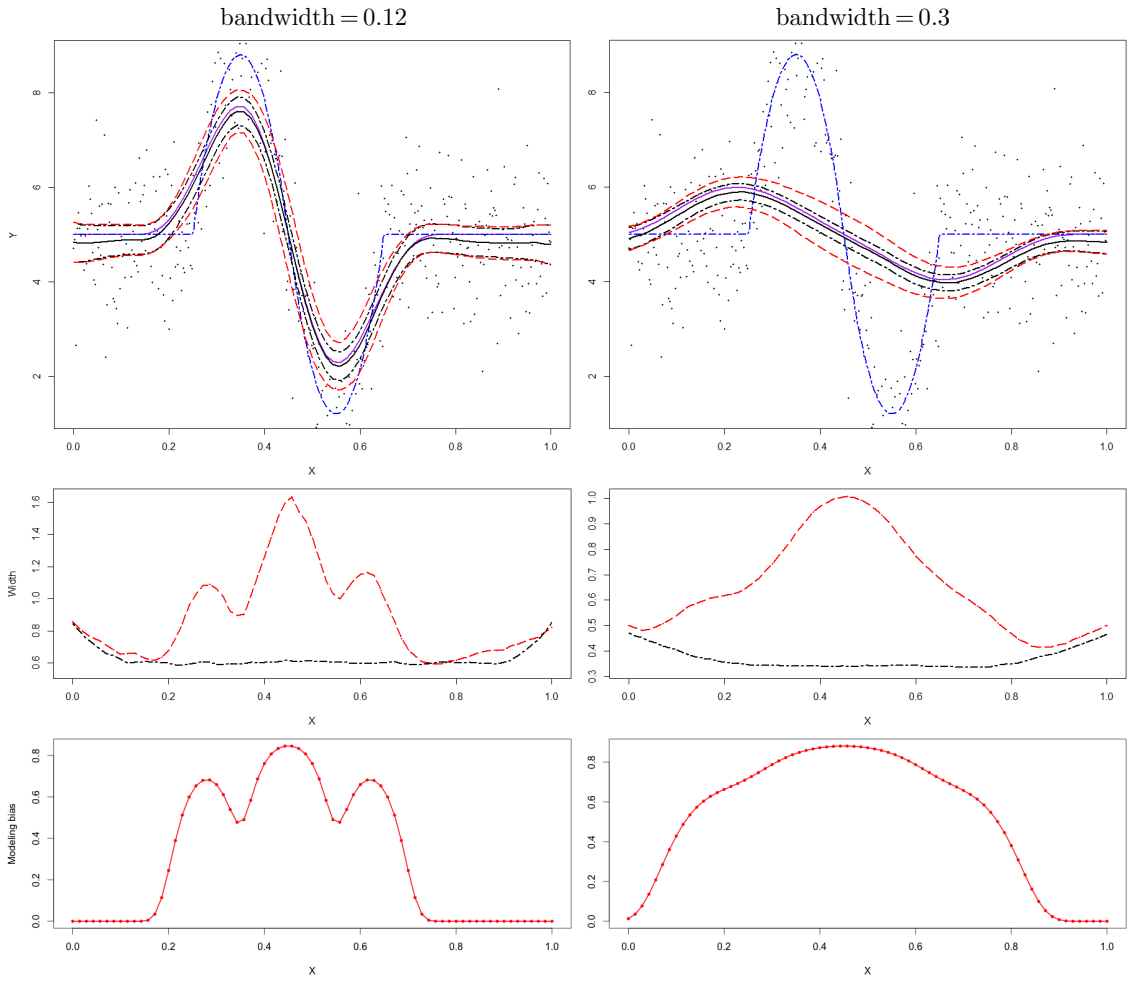
Table 3.1: Effective coverage probabilities for the local constant regression

	Confidence levels									
h	**0.95**	**0.90**	**0.85**	**0.80**	**0.75**	**0.70**	**0.65**	**0.60**	**0.55**	**0.50**
0.12	0.971	0.947	0.917	0.888	0.863	0.830	0.800	0.769	0.738	0.702
0.3	0.982	0.963	0.942	0.918	0.895	0.868	0.842	0.815	0.784	0.750

3.3.6 Correction for multiplicity

Here we compare the $\boldsymbol{Y}$ and the bootstrap corrections for multiplicity, i.e. the values $\mathfrak{c}(\alpha)$ and $\mathfrak{c}^\circ(\alpha)$ defined in (1.18) and (3.4). The numerical results in Tables 3.2, 3.3

Figure 3.1: **Local constant regression:**
Confidence bands, their widths, and the modelling bias



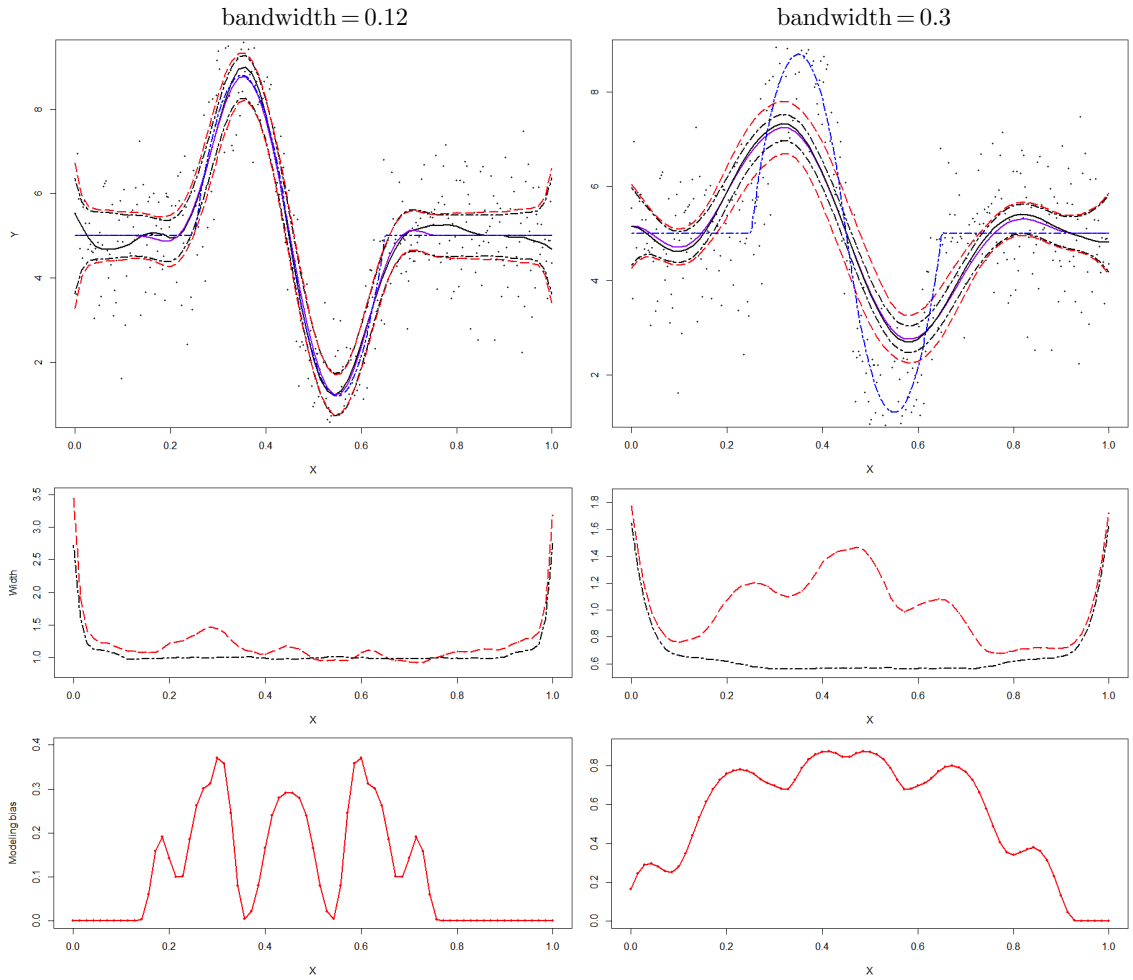
Legend for the top graphs:

- 90% bootstrap simultaneous confidence band
- the true function $f(x)$
- 90% MC simultaneous confidence band
- local constant MLE
- smoothed target function

Legend for the middle and the bottom graphs:

- width of the 90% bootstrap confidence bands from the upper graphs
- width of the 90% MC confidence bands from the upper graphs
- modelling bias from the expression (3.14)

Figure 3.2: **Local quadratic regression:**
Confidence bands, their widths, and the modelling bias



Legend for the top graphs:

- 90% bootstrap simultaneous confidence band
- the true function $f(x)$
- 90% MC simultaneous confidence band
- local constant MLE
- smoothed target function

Legend for the middle and the bottom graphs:

- width of the 90% bootstrap confidence bands from the upper graphs
- width of the 90% MC confidence bands from the upper graphs
- modelling bias from the expression (3.15)

are based on 10^4 $\{Y_i\} \sim (3.12)$ independent samples and 10^4 independent bootstrap samples $\{u_i\} \sim \mathcal{N}(1,1)$. The second line in Tables 3.2, 3.3 contains the range of the nominal confidence levels $(1-\alpha) = 0.95, 0.9, \ldots, 0.5$ (similarly to the Table 3.1). The first column contains the values of the bandwidth $h = 0.12, 0.3$, and the second column – the resampling scheme: Monte Carlo (MC) or bootstrap (B). The Monte Carlo experiment yields the corrected confidence levels $1 - \mathfrak{c}(\alpha)$, and the bootstrap yields $1 - \mathfrak{c}^{\circ}(\alpha)$. The lines 3–6 contain the average values of $1 - \mathfrak{c}(\alpha)$ and $1 - \mathfrak{c}^{\circ}(\alpha)$ over all the experiments. The results show that for the smaller bandwidth both the MC and bootstrap corrections are bigger than the ones for the larger bandwidth. In the case of a smaller bandwidth the local models have less intersections with each other, and hence, the corrections for multiplicity are closer to the Bonferroni's bound.

Table 3.2: **Local constant regression:**
MC vs Bootstrap confidence levels corrected for multiplicity

		Confidence levels									
h	r.m.	**0.95**	**0.90**	**0.85**	**0.80**	**0.75**	**0.70**	**0.65**	**0.60**	**0.55**	**0.50**
0.12	MC	0.997	0.994	0.989	0.985	0.980	0.975	0.969	0.963	0.956	0.949
	B	0.998	0.995	0.991	0.988	0.984	0.979	0.975	0.969	0.963	0.957
0.3	MC	0.993	0.983	0.973	0.962	0.949	0.936	0.922	0.906	0.891	0.873
	B	0.994	0.986	0.977	0.968	0.958	0.947	0.935	0.922	0.908	0.893

Table 3.3: **Local quadratic regression:**
MC vs Bootstrap confidence levels corrected for multiplicity

		Confidence levels									
h	r.m.	**0.95**	**0.90**	**0.85**	**0.80**	**0.75**	**0.70**	**0.65**	**0.60**	**0.55**	**0.50**
0.12	MC	0.997	0.993	0.989	0.985	0.979	0.974	0.968	0.961	0.954	0.946
	B	0.998	0.995	0.991	0.988	0.984	0.979	0.974	0.969	0.963	0.956
0.3	MC	0.993	0.983	0.973	0.961	0.949	0.936	0.921	0.904	0.887	0.868
	B	0.996	0.991	0.985	0.978	0.971	0.963	0.954	0.944	0.934	0.923

Remark 3.3. The theoretical results of this chapter can be extended to the case when a set of considered local models has cardinality of the continuum, and the confidence bands are uniform w.r.t. the local parameter. This extension would require some

uniform statements such as locally uniform square-root Wilks approximation (see e.g. Spokoiny and Zhilova (2013)).

Remark 3.4. The use of the bootstrap procedure in the problem of choosing an optimal bandwidth is considered in Spokoiny and Willrich (2015).

3.4 Conditions

Here we show conditions required for the main results. The conditions in Section 3.4.1 come from the general finite sample theory by Spokoiny (2012a), they are required for the results of Sections A.1 and A.2. The conditions in Section 3.4.2 are necessary to prove the statements on multiplier bootstrap validity.

3.4.1 Basic conditions

Introduce the stochastic part of the k-th likelihood process: $\zeta_k(\boldsymbol{\theta}) \stackrel{\text{def}}{=} L_k(\boldsymbol{\theta}) - I\!\!E L_k(\boldsymbol{\theta})$, and its marginal summand: $\zeta_{i,k}(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \ell_{i,k}(\boldsymbol{\theta}) - I\!\!E \ell_{i,k}(\boldsymbol{\theta})$ for $\ell_{i,k}(\boldsymbol{\theta})$ defined in (3.1).

($\boldsymbol{ED_{0,k}}$) *For each $k = 1, \ldots, K$ there exist a positive-definite $p_k \times p_k$ symmetric matrix V_k^2 and constants $\mathtt{g}_k > 0, \nu_k \geq 1$ such that* $\operatorname{Var}\{\nabla_{\boldsymbol{\theta}} \zeta_k(\boldsymbol{\theta}_k^*)\} \leq V_k^2$ *and*

$$\sup_{\boldsymbol{\gamma} \in I\!R^{p_k}} \log I\!\!E \exp\left\{ \lambda \frac{\boldsymbol{\gamma}^\top \nabla_{\boldsymbol{\theta}} \zeta_k(\boldsymbol{\theta}_k^*)}{\|V_k \boldsymbol{\gamma}\|} \right\} \leq \nu_k^2 \lambda^2 / 2, \qquad |\lambda| \leq \mathtt{g}_k.$$

($\boldsymbol{ED_{2,k}}$) *For each $k = 1, \ldots, K$ there exist a constant $\omega_k > 0$ and for each $\mathtt{r} > 0$ a constant $\mathtt{g}_{2,k}(\mathtt{r})$ such that it holds for all $\boldsymbol{\theta} \in \Theta_{0,k}(\mathtt{r})$ and for $j = 1, 2$*

$$\sup_{\substack{\boldsymbol{\gamma}_j \in I\!R^{p_k} \\ \|\boldsymbol{\gamma}_j\| \leq 1}} \log I\!\!E \exp\left\{ \frac{\lambda}{\omega_k} \boldsymbol{\gamma}_1^\top D_k^{-1} \nabla^2_{\boldsymbol{\theta}} \zeta_k(\boldsymbol{\theta}) D_k^{-1} \boldsymbol{\gamma}_2 \right\} \leq \nu_k^2 \lambda^2 / 2, \qquad |\lambda| \leq \mathtt{g}_{2,k}(\mathtt{r}).$$

($\boldsymbol{\mathcal{L}_{0,k}}$) *For each $k = 1, \ldots, K$ and for each $\mathtt{r} > 0$ there exists a constant $\delta_k(\mathtt{r}) \geq 0$ such that for $\mathtt{r} \leq \mathtt{r}_{0,k}$ ($\mathtt{r}_{0,k}$ come from condition* (A.2) *of Theorem A.1 in Section A.1) $\delta_k(\mathtt{r}) \leq 1/2$, and for all $\boldsymbol{\theta} \in \Theta_{0,k}(\mathtt{r})$ it holds*

$$\|D_k^{-1} \check{D}_k^2(\boldsymbol{\theta}) D_k^{-1} - \boldsymbol{I}_{p_k}\| \leq \delta_k(\mathtt{r}),$$

where $\check{D}_k^2(\boldsymbol{\theta}) \stackrel{\text{def}}{=} -\nabla^2_{\boldsymbol{\theta}} I\!\!E L_k(\boldsymbol{\theta})$ and $\Theta_{0,k}(\mathtt{r}) \stackrel{\text{def}}{=} \{\boldsymbol{\theta} \in \Theta_k : \|D_k(\boldsymbol{\theta} - \boldsymbol{\theta}_k^*)\| \leq \mathtt{r}\}$.

($\boldsymbol{\mathcal{I}_k}$) *There exist constants $\mathfrak{a}_k > 0$ for all $k = 1, \ldots, K$ s.t.*

$$\mathfrak{a}_k^2 D_k^2 \geq V_k^2.$$

Denote $\widehat{\mathfrak{a}}^2 \stackrel{\text{def}}{=} \max_{1 \leq k \leq K} \mathfrak{a}_k^2$.

($\mathcal{L}\mathtt{r}_k$) *For each* $k = 1, \ldots, K$ *and* $\mathtt{r} \geq \mathtt{r}_{0,k}$ *there exists a value* $\mathtt{b}_k(\mathtt{r}) > 0$ *s.t.* $\mathtt{rb}_k(\mathtt{r}) \to \infty$ *for* $\mathtt{r} \to \infty$ *and* $\forall \boldsymbol{\theta} \in \Theta_k : \|D_k(\boldsymbol{\theta} - \boldsymbol{\theta}_k^*)\| = \mathtt{r}$ *it holds*

$$-2\{I\!\!E L_k(\boldsymbol{\theta}) - I\!\!E L_k(\boldsymbol{\theta}_k^*)\} \geq \mathtt{r}^2 \mathtt{b}_k(\mathtt{r}).$$

3.4.2 Conditions required for the bootstrap validity

($\widehat{\mathbf{SmB}}$) *There exists a constant* $\widehat{\delta}_{\mathrm{smb}} \geq 0$ *such that it holds for the matrices* B_k^2 *and* H_k^2 *defined in* (3.10)*:*

$$\max_{1 \leq k \leq K} \left\| H_k^{-1} B_k^2 H_k^{-1} \right\| \leq \widehat{\delta}_{\mathrm{smb}}^2,$$
$$\widehat{\delta}_{\mathrm{smb}}^2 \leq \mathtt{C} \left(\frac{n}{p_{\max}^{13}} \right)^{1/8} \log^{-7/8}(K) \log^{-3/8}(n p_{\mathrm{sum}}).$$

($\boldsymbol{ED}_{\mathbf{2}\boldsymbol{m},\boldsymbol{k}}$) *For each* $k = 1, \ldots, K$, $\mathtt{r} > 0$, $i = 1, \ldots, n$, $j = 1, 2$ *and for all* $\boldsymbol{\theta} \in \Theta_{0,k}(\mathtt{r})$ *it holds for the values* $\omega_k \geq 0$ *and* $\mathtt{g}_{2,k}(\mathtt{r})$ *from the condition* **($\boldsymbol{ED}_{\mathbf{2},\boldsymbol{k}}$)***:*

$$\sup_{\substack{\boldsymbol{\gamma}_j \in I\!R^{p_k} \\ \|\boldsymbol{\gamma}_j\| \leq 1}} \log I\!\!E \exp\left\{ \frac{\lambda}{\omega_k} \boldsymbol{\gamma}_1^\top D_k^{-1} \nabla_{\boldsymbol{\theta}}^2 \zeta_{i,k}(\boldsymbol{\theta}) D_k^{-1} \boldsymbol{\gamma}_2 \right\} \leq \frac{\nu_0^2 \lambda^2}{2n}, \qquad |\lambda| \leq \mathtt{g}_{2,k}(\mathtt{r}),$$

($\mathcal{L}_{\mathbf{0}\boldsymbol{m},\boldsymbol{k}}$) *For each* $k = 1, \ldots, K$, $\mathtt{r} > 0$, $i = 1, \ldots, n$ *and for all* $\boldsymbol{\theta} \in \Theta_{0,k}(\mathtt{r})$ *there exists a value* $\mathtt{C}_{m,k}(\mathtt{r}) \geq 0$ *such that*

$$\|D_k^{-1} \nabla_{\boldsymbol{\theta}}^2 I\!\!E \ell_{i,k}(\boldsymbol{\theta}) D_k^{-1}\| \leq \mathtt{C}_{m,k}(\mathtt{r}) n^{-1}.$$

($\mathcal{I}_{\boldsymbol{B},\boldsymbol{k}}$) *For each* $k = 1, \ldots, K$ *there exists a constant* $\mathfrak{a}_{B,k}^2 > 0$ *s.t.*

$$\mathfrak{a}_{B,k}^2 D_k^2 \geq B_k^2.$$

Denote $\widehat{\mathfrak{a}}_B^2 \stackrel{\mathrm{def}}{=} \max_{1 \leq k \leq K} \mathfrak{a}_{B,k}^2$.

($\widehat{\boldsymbol{SD}_{\mathbf{1}}}$) *There exists a constant* $0 \leq \delta_{v^*}^2 \leq \mathtt{C} p_{\mathrm{sum}}/n$ *such that it holds for all* $i = 1, \ldots, n$ *with exponentially high probability*

$$\left\| \widehat{H}^{-1} \left\{ \boldsymbol{g}_i \boldsymbol{g}_i^\top - I\!\!E \left[\boldsymbol{g}_i \boldsymbol{g}_i^\top \right] \right\} \widehat{H}^{-1} \right\| \leq \delta_{v^*}^2,$$

where

$$\boldsymbol{g}_i \stackrel{\mathrm{def}}{=} \left(\nabla_{\boldsymbol{\theta}} \ell_{i,1}(\boldsymbol{\theta}_1^*)^\top, \ldots, \nabla_{\boldsymbol{\theta}} \ell_{i,K}(\boldsymbol{\theta}_K^*)^\top \right)^\top \in I\!R^{p_{\mathrm{sum}}},$$
$$\widehat{H}^2 \stackrel{\mathrm{def}}{=} \sum\nolimits_{i=1}^n I\!\!E \left\{ \boldsymbol{g}_i \boldsymbol{g}_i^\top \right\},$$
$$p_{\mathrm{sum}} \stackrel{\mathrm{def}}{=} p_1 + \cdots + p_K.$$

(*Eb*) *The i.i.d. bootstrap weights* u_i *are independent of* $\boldsymbol{Y}$ *, and for all* $i = 1, \ldots, n$ *it holds for some constants* $\mathsf{g} > 0, \nu \geq 1$

$$\mathbb{E} u_i = 1, \quad \operatorname{Var} u_i = 1,$$
$$\log \mathbb{E} \exp\{\lambda(u_i - 1)\} \leq \nu_0^2 \lambda^2 / 2, \qquad |\lambda| \leq \mathsf{g}.$$

Appendix A

Square-root Wilks approximations

This chapter considers the non-asymptotic Wilks approximation theorem. Section A.1 restates the results by Spokoiny (2012a, 2013). In Section A.2 we prove the Wilks theorem for the bootstrap world. In Section A.3 these results are specified for some common models: i.i.d. observations, generalised linear model and linear median regression, we also show the dependence of the non-asymptotic bounds on sample size and parameter's dimension.

A.1 Finite sample theory

Let us use the notations given in the introduction: $L(\boldsymbol{\theta})$ is the log-likelihood process, which depends on the data $\boldsymbol{Y}$ and corresponds to the parametric family of probability distributions $\{I\!P_{\boldsymbol{\theta}}\}$. The general finite sample approach by Spokoiny (2012a) does not require the true measure $I\!P$ to belong to $\{I\!P_{\boldsymbol{\theta}}\}$. The target parameter $\boldsymbol{\theta}^*$ is defined as in (1.3) by projection of the true measure $I\!P$ on $\{I\!P_{\boldsymbol{\theta}}\}$. D_0^2 denotes the full Fisher information $p \times p$ matrix, which is deterministic, symmetric and positive-definite:

$$D_0^2 \stackrel{\text{def}}{=} -\nabla^2_{\boldsymbol{\theta}} I\!E L(\boldsymbol{\theta}^*).$$

A centered p-dimensional random vector $\boldsymbol{\xi}$ denotes the normalised score:

$$\boldsymbol{\xi} \stackrel{\text{def}}{=} D_0^{-1} \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}^*).$$

Introduce the following elliptic vicinity around the true point $\boldsymbol{\theta}^*$:

$$\Theta_0(\mathtt{r}) \stackrel{\text{def}}{=} \{\boldsymbol{\theta} : \|D_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \le \mathtt{r}\}. \tag{A.1}$$

The non-asymptotic Wilks approximating bound by Spokoiny (2012a), Spokoiny (2013) requires that the maximum likelihood estimate $\widetilde{\boldsymbol{\theta}}$ gets into the local vicinity $\Theta_0(\mathtt{r}_0)$ of some radius $\mathtt{r}_0 > 0$ with probability $\geq 1 - 3\mathrm{e}^{-\mathtt{x}}$, $\mathtt{x} > 0$. This is guaranteed by the following concentration result:

Theorem A.1 (Concentration of MLE, Spokoiny (2013)). *Let the conditions* $(\boldsymbol{ED_0})$, $(\boldsymbol{ED_2})$, $(\boldsymbol{\mathcal{I}})$ *and* $(\boldsymbol{\mathcal{L}\mathtt{r}})$ *be fulfilled. If for the constant* $\mathtt{r}_0 > 0$ *and for the function* $\mathtt{b}(\mathtt{r})$ *from* $(\boldsymbol{\mathcal{L}\mathtt{r}})$*:*

$$\mathtt{b}(\mathtt{r})\mathtt{r} \geq 2\left\{ \mathfrak{Z}_{\mathtt{qf}}(\mathtt{x}, I\!B) + 6\omega\nu_0\, \mathfrak{Z}(\mathtt{x} + \log(2\mathtt{r}/\mathtt{r}_0)) \right\}, \quad \mathtt{r} > \mathtt{r}_0 \tag{A.2}$$

where the functions $\mathfrak{Z}(\mathtt{x})$ *and* $\mathfrak{Z}_{\mathtt{qf}}(\mathtt{x}, I\!B)$ *are defined respectively in* (A.5) *and* (A.6)*, then it holds*

$$I\!P\left(\widetilde{\boldsymbol{\theta}} \notin \Theta_0(\mathtt{r}_0)\right) \leq 3\mathrm{e}^{-\mathtt{x}}.$$

The constants ω, ν_0 *and* $\mathfrak{a}$ *come from the imposed conditions* $(\boldsymbol{ED_0})$–$(\boldsymbol{\mathcal{I}})$ *(from Section 2.5). In the case A.3.1* $\mathtt{r}_0 \geq \mathtt{C}\sqrt{p + \mathtt{x}}$.

The following result is one of the central in the general finite sample theory and is crucial for the present study due to the scheme (1.6):

Theorem A.2 (Wilks approximation, Spokoiny (2013)). *Under the conditions of Theorem A.1 for some* $\mathtt{r}_0 > 0$ *s.t.* (A.2) *is fulfilled, and under condition* $(\boldsymbol{\mathcal{L}_0})$ *it holds with probability* $\geq 1 - 5\mathrm{e}^{-\mathtt{x}}$

$$\left| 2\Big\{ L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*) \Big\} - \|\boldsymbol{\xi}\|^2 \right| \leq \Delta_{\mathrm{W}^2}(\mathtt{r}_0, \mathtt{x}),$$
$$\left| \sqrt{2\Big\{ L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*) \Big\}} - \|\boldsymbol{\xi}\| \right| \leq \Delta_{\mathrm{W}}(\mathtt{r}_0, \mathtt{x})$$

for

$$\Delta_{\mathrm{W}}(\mathtt{r}, \mathtt{x}) \stackrel{\mathrm{def}}{=} 3\mathtt{r}\left\{ \delta(\mathtt{r}) + 6\nu_0\, \mathfrak{Z}(\mathtt{x})\omega \right\}, \tag{A.3}$$

$$\Delta_{\mathrm{W}^2}(\mathtt{r}, \mathtt{x}) \stackrel{\mathrm{def}}{=} \frac{2}{3}\left\{ 2\mathtt{r} + \mathfrak{Z}_{\mathtt{qf}}(\mathtt{x}, I\!B) \right\} \Delta_{\mathrm{W}}(\mathtt{r}, \mathtt{x}), \tag{A.4}$$

$$\mathfrak{Z}(\mathtt{x}) \stackrel{\mathrm{def}}{=} 2\sqrt{p} + \sqrt{2\mathtt{x}} + 4p(\mathtt{x}\mathtt{g}^{-2} + 1)\mathtt{g}^{-1}. \tag{A.5}$$

In the case A.3.1 it holds for $\mathtt{r} \leq \mathtt{r}_0$*:*

$$\Delta_{\mathrm{W}}(\mathtt{r}, \mathtt{x}) \leq \mathtt{C}\,\frac{p + \mathtt{x}}{\sqrt{n}}, \quad \Delta_{\mathrm{W}^2}(\mathtt{r}, \mathtt{x}) \leq \mathtt{C}\sqrt{\frac{(p + \mathtt{x})^3}{n}}.$$

The constants $\mathtt{g}$ *and* $\delta(\mathtt{r})$ *come from the imposed conditions* $(\boldsymbol{ED_0})$, $(\boldsymbol{\mathcal{L}_0})$ *(from Section 2.5), and the function* $\mathfrak{Z}_{\mathtt{qf}}(\mathtt{x}, I\!B)$*, defined in* (A.6)*, corresponds to the quantile function of deviations of the random value* $\|\boldsymbol{\xi}\|$ *(see Theorem A.3 below).*

The following theorem characterizes the tail behaviour of the approximating term $\|\boldsymbol{\xi}\|^2$. It means that with a bounded exponential moment of the vector $\boldsymbol{\xi}$ (condition $(\boldsymbol{ED_0})$) its squared Euclidean norm $\|\boldsymbol{\xi}\|^2$ has three regimes of deviations: sub-Gaussian, Poissonian and large-deviations' zone.

Theorem A.3 (Deviation bound for a random quadratic form, Spokoiny (2012b))**.** *Let condition* $(\boldsymbol{ED_0})$ *be fulfilled, then for* $\mathtt{g} \geq \sqrt{2\,\mathrm{tr}(I\!B^2)}$ *it holds:*

$$I\!P\left(\|\boldsymbol{\xi}\|^2 \geq \mathfrak{Z}^2_{\mathtt{qf}}(\mathtt{x}, I\!B)\right) \leq 2\mathrm{e}^{-\mathtt{x}} + 8.4\mathrm{e}^{-\mathtt{x}_c},$$

where $I\!B^2 \stackrel{\mathrm{def}}{=} D_0^{-1} V_0^2 D_0^{-1}$, $\lambda(I\!B)$ *is a maximum eigenvalue of* $I\!B^2$,

$$\mathfrak{Z}^2_{\mathtt{qf}}(\mathtt{x}, I\!B) \stackrel{\mathrm{def}}{=} \begin{cases} \mathrm{tr}(I\!B^2) + \sqrt{8\,\mathrm{tr}(I\!B^4)\mathtt{x}}, & \mathtt{x} \leq \sqrt{2\,\mathrm{tr}(I\!B^4)}/\{18\lambda(I\!B)\}, \\ \mathrm{tr}(I\!B^2) + 6\mathtt{x}\lambda(I\!B), & \sqrt{2\,\mathrm{tr}(I\!B^4)}/\{18\lambda(I\!B)\} < \mathtt{x} \leq \mathtt{x}_c, \\ |\mathtt{z}_c + 2(\mathtt{x} - \mathtt{x}_c)/\mathtt{g}_c|^2\,\lambda(I\!B), & \mathtt{x} > \mathtt{x}_c, \end{cases} \tag{A.6}$$

$$\begin{aligned} 2\mathtt{x}_c &\stackrel{\mathrm{def}}{=} 2\mathtt{x}_c(I\!B) \stackrel{\mathrm{def}}{=} \mu_c \mathtt{z}_c{}^2 + \log\det\left(\boldsymbol{I}_p - \mu_c I\!B^2/\lambda(I\!B)\right), \\ \mathtt{z}_c{}^2 &\stackrel{\mathrm{def}}{=} \left\{\mathtt{g}^2/\mu_c^2 - \mathrm{tr}\left(I\!B^2\right)/\mu_c\right\}/\lambda(I\!B), \\ \mathtt{g}_c &\stackrel{\mathrm{def}}{=} \sqrt{\mathtt{g}^2 - \mu_c\,\mathrm{tr}\left(I\!B^2\right)}/\sqrt{\lambda(I\!B)}, \\ \mu_c &\stackrel{\mathrm{def}}{=} 2/3. \end{aligned} \tag{A.7}$$

The matrix V_0^2 *comes from condition* $(\boldsymbol{ED_0})$ *and can be defined as*

$$V_0^2 \stackrel{\mathrm{def}}{=} \mathrm{Var}\left\{\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}^*)\right\}.$$

By condition $(\boldsymbol{\mathcal{I}})$ $\mathrm{tr}(I\!B^2) \leq \mathfrak{a}^2 p$, $\mathrm{tr}(I\!B^4) \leq \mathfrak{a}^4 p$ *and* $\lambda(I\!B) \leq \mathfrak{a}^2$. *In the case A.3.1* $\mathtt{g} = \mathtt{C}\sqrt{n}$, *hence* $\mathtt{x}_c = \mathtt{C}n$, *and for* $\mathtt{x} \leq \mathtt{x}_c$ *it holds:*

$$\mathfrak{Z}^2_{\mathtt{qf}}(\mathtt{x}, I\!B) \leq \mathtt{C}\mathfrak{a}^2(p + 6\mathtt{x}). \tag{A.8}$$

A.2 Finite sample theory for the bootstrap world

Let us introduce the bootstrap score vector at a point $\boldsymbol{\theta} \in \Theta$:

$$\begin{aligned} \boldsymbol{\xi}^\circ(\boldsymbol{\theta}) &\stackrel{\mathrm{def}}{=} D_0^{-1} \nabla_{\boldsymbol{\theta}} \boldsymbol{\zeta}^\circ(\boldsymbol{\theta}) \\ &= \sum\nolimits_{i=1}^n D_0^{-1} \nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta})(u_i - 1). \end{aligned} \tag{A.9}$$

Theorem A.4 (Bootstrap Wilks approximation)**.** *Under the conditions of Theorems A.1 and A.5 for some* $\mathtt{r}_0{}^2 \geq 0$ *s.t.* (A.2) *and* (A.45) *are fulfilled, it holds with* $I\!P$ *-probability* $\geq 1 - 5\mathrm{e}^{-\mathtt{x}}$

$$I\!P^\circ\left(\left|\sup_{\boldsymbol{\theta}\in\Theta} 2\left\{L^\circ(\boldsymbol{\theta}) - L^\circ(\widetilde{\boldsymbol{\theta}})\right\} - \|\boldsymbol{\xi}^\circ(\widetilde{\boldsymbol{\theta}})\|^2\right| \leq \Delta^\circ_{W^2}(\mathtt{r}_0,\mathtt{x})\right) \geq 1 - 4\mathrm{e}^{-\mathtt{x}},$$
$$I\!P^\circ\left(\left|\sqrt{\sup_{\boldsymbol{\theta}\in\Theta} 2\left\{L^\circ(\boldsymbol{\theta}) - L^\circ(\widetilde{\boldsymbol{\theta}})\right\}} - \|\boldsymbol{\xi}^\circ(\widetilde{\boldsymbol{\theta}})\|\right| \leq \Delta^\circ_{W}(\mathtt{r}_0,\mathtt{x})\right) \geq 1 - 4\mathrm{e}^{-\mathtt{x}}.$$

where the error terms $\Delta^\circ_W(\mathtt{r},\mathtt{x}), \Delta^\circ_{W^2}(\mathtt{r},\mathtt{x})$ *are deterministic and*

$$\begin{aligned} \Delta^\circ_W(\mathtt{r},\mathtt{x}) &\stackrel{\text{def}}{=} 2\Delta_{\mathrm{W}}(\mathtt{r},\mathtt{x}) + 36\nu_0\mathtt{r}\omega_1(\mathtt{r})\,\mathfrak{Z}(\mathtt{x}), \\ \Delta^\circ_{W^2}(\mathtt{r},\mathtt{x}) &\stackrel{\text{def}}{=} \frac{1}{18}\left\{12\mathtt{r}\Delta^\circ_W(\mathtt{r},\mathtt{x}) + \Delta^\circ_W(\mathtt{r},\mathtt{x})^2\right\}. \end{aligned} \tag{A.10}$$

$\Delta_{\mathrm{W}}(\mathtt{r},\mathtt{x})$ *and* $\mathfrak{Z}(\mathtt{x})$ *are defined respectively in* (A.3) *and* (A.5)*, and* $\omega_1(\mathtt{r})$ *is given in* (A.18)*. For the case A.3.1 and* $\mathtt{r} \leq \mathtt{r}_0$ *it holds:*

$$\Delta^\circ_W(\mathtt{r},\mathtt{x}) \leq \mathtt{C}\,\frac{p+\mathtt{x}}{\sqrt{n}}\sqrt{\mathtt{x}}, \quad \Delta^\circ_{W^2}(\mathtt{r},\mathtt{x}) \leq \mathtt{C}\sqrt{\frac{(p+\mathtt{x})^3}{n}}\sqrt{\mathtt{x}}.$$

Proof of Theorem A.4. Let us consider the following random process in the bootstrap world for $\boldsymbol{\theta},\boldsymbol{\theta}_1 \in \Theta_0(\mathtt{r})$:

$$\mathcal{A}^\circ(\boldsymbol{\theta},\boldsymbol{\theta}_1) \stackrel{\text{def}}{=} L^\circ(\boldsymbol{\theta}) - L^\circ(\boldsymbol{\theta}_1) - (\boldsymbol{\theta}-\boldsymbol{\theta}_1)^\top\nabla_{\boldsymbol{\theta}}L^\circ(\boldsymbol{\theta}_1) + \frac{1}{2}\|D_0(\boldsymbol{\theta}-\boldsymbol{\theta}_1)\|^2.$$

It holds $\mathcal{A}^\circ(\boldsymbol{\theta}_1,\boldsymbol{\theta}_1) = 0$. Taylor expansion w.r.t. $\boldsymbol{\theta}$ around $\boldsymbol{\theta}_1$ implies :

$$\mathcal{A}^\circ(\boldsymbol{\theta},\boldsymbol{\theta}_1) = (\boldsymbol{\theta}-\boldsymbol{\theta}_1)^\top\nabla_{\boldsymbol{\theta}}\mathcal{A}^\circ(\overline{\boldsymbol{\theta}}_1,\boldsymbol{\theta}_1),$$

where $\overline{\boldsymbol{\theta}}_1$ is some convex combination of the vectors $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_1$. Therefore,

$$|\mathcal{A}^\circ(\boldsymbol{\theta},\boldsymbol{\theta}_1)| \leq \|D_0(\boldsymbol{\theta}-\boldsymbol{\theta}_1)\| \sup_{\boldsymbol{\theta}\in\Theta_0(\mathtt{r})} \left\|D_0^{-1}\nabla_{\boldsymbol{\theta}}\mathcal{A}^\circ(\boldsymbol{\theta},\boldsymbol{\theta}_1)\right\| \tag{A.11}$$
$$\leq 2\mathtt{r} \sup_{\boldsymbol{\theta}\in\Theta_0(\mathtt{r})} \left\|D_0^{-1}\nabla_{\boldsymbol{\theta}}\mathcal{A}^\circ(\boldsymbol{\theta},\boldsymbol{\theta}_1)\right\|. \tag{A.12}$$

Now let us consider the normalized gradient process:

$$D_0^{-1}\nabla_{\boldsymbol{\theta}}\mathcal{A}^\circ(\boldsymbol{\theta},\boldsymbol{\theta}_1) = D_0^{-1}\left\{\nabla_{\boldsymbol{\theta}}L^\circ(\boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}}L^\circ(\boldsymbol{\theta}_1)\right\} + D_0(\boldsymbol{\theta}-\boldsymbol{\theta}_1).$$

The deterministic part of it reads as:

$$D_0^{-1}\nabla_{\boldsymbol{\theta}}I\!E^\circ\mathcal{A}^\circ(\boldsymbol{\theta},\boldsymbol{\theta}_1) = D_0^{-1}\left\{\nabla_{\boldsymbol{\theta}}L(\boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}}L(\boldsymbol{\theta}_1)\right\} + D_0(\boldsymbol{\theta}-\boldsymbol{\theta}_1).$$

Proposition 3.1 in Spokoiny (2013) implies due to the conditions $(\mathcal{L}_0)$, (ED_2), that the following random event holds with $I\!P$-probability at least $1 - \mathrm{e}^{-\mathtt{x}}$ for all $\boldsymbol{\theta}, \boldsymbol{\theta}_1 \in \Theta_0(\mathtt{r})$ and $\mathtt{r} \le \mathtt{r}_0$:

$$\begin{aligned} \| D_0^{-1} \nabla_{\boldsymbol{\theta}} I\!E^{\circ} \mathcal{A}^{\circ}(\boldsymbol{\theta}, \boldsymbol{\theta}_1) \| &= \left\| D_0^{-1} \left\{ \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}_1) \right\} + D_0 (\boldsymbol{\theta} - \boldsymbol{\theta}_1) \right\| \\ &\le \frac{2}{3} \Delta_{\mathrm{W}}(\mathtt{r}, \mathtt{x}), \end{aligned} \tag{A.13}$$

where the deterministic error term $\Delta_{\mathrm{W}}(\mathtt{r}, \mathtt{x})$ is given in (A.3).

Denote the stochastic part of $D_0^{-1} \nabla_{\boldsymbol{\theta}} \mathcal{A}^{\circ}(\boldsymbol{\theta}, \boldsymbol{\theta}_1)$ as follows:

$$\begin{aligned} \mathcal{Y}^{\circ}(\boldsymbol{\theta}, \boldsymbol{\theta}_1) &\stackrel{\mathrm{def}}{=} D_0^{-1} \left\{ \nabla_{\boldsymbol{\theta}} \mathcal{A}^{\circ}(\boldsymbol{\theta}, \boldsymbol{\theta}_1) - \nabla_{\boldsymbol{\theta}} I\!E^{\circ} \mathcal{A}^{\circ}(\boldsymbol{\theta}, \boldsymbol{\theta}_1) \right\} \\ &= \sum\nolimits_{i=1}^{n} D_0^{-1} \left\{ \nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}_1) \right\} (u_i - 1). \end{aligned}$$

In order to bound its norm's supremum w.r.t. $\boldsymbol{\theta} \in \Theta_0(\mathtt{r})$ for $\mathtt{r} \le \mathtt{r}_0$ we use the idea from the proof of Proposition 3.1 in Spokoiny (2013). Let us introduce the new parameters $\boldsymbol{\upsilon} \stackrel{\mathrm{def}}{=} D_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$ and $\boldsymbol{\upsilon}_1 \stackrel{\mathrm{def}}{=} D_0(\boldsymbol{\theta}_1 - \boldsymbol{\theta}^*)$, then

$$\nabla_{\boldsymbol{\upsilon}} \mathcal{Y}^{\circ}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}_1) = \sum\nolimits_{i=1}^{n} D_0^{-1} \nabla_{\boldsymbol{\theta}}^2 \ell_i(\boldsymbol{\theta}) D_0^{-1} (u_i - 1).$$

Thus, we obtain a proper normalisation for $\nabla_{\boldsymbol{\upsilon}} \mathcal{Y}^{\circ}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}_1)$. Independence of $u_1, \ldots, u_n$ and Lemma A.1 imply with probability $\ge 1 - \mathrm{e}^{-\mathtt{x}}$ for $j = 1, 2$ and $\omega_1(\mathtt{r})$ given in (A.18):

$$\sup_{\substack{\boldsymbol{\gamma}_j \in I\!R^p \\ \|\boldsymbol{\gamma}_j\| = 1}} \log I\!E^{\circ} \exp \left\{ \frac{\lambda}{\omega_1(\mathtt{r})} \boldsymbol{\gamma}_1^{\top} \nabla_{\boldsymbol{\upsilon}} \mathcal{Y}^{\circ}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}_1) \boldsymbol{\gamma}_2 \right\} \le \frac{\lambda^2 \nu_0^2}{2}, \quad |\lambda| \le \mathtt{g}_2(\mathtt{r}).$$

This allows to apply Theorem A.3 from Spokoiny (2013) on a uniform bound for the norm of stochastic process to $\omega_1^{-1}(\mathtt{r}) \mathcal{Y}^{\circ}(\boldsymbol{\theta}, \boldsymbol{\theta}_1)$. By the triangle inequality it holds for $\mathtt{r} \le \mathtt{r}_0$:

$$I\!P^{\circ} \left(\sup_{\boldsymbol{\theta}, \boldsymbol{\theta}_1 \in \Theta_0(\mathtt{r})} \| \mathcal{Y}^{\circ}(\boldsymbol{\theta}, \boldsymbol{\theta}_1) \| \le 12 \nu_0 \mathtt{r} \omega_1(\mathtt{r}) \, \mathfrak{Z}(\mathtt{x}) \right) \ge 1 - \mathrm{e}^{-\mathtt{x}}, \tag{A.14}$$

where $\mathfrak{Z}(\mathtt{x})$ is defined in (A.5). Collecting together the bounds (A.12), (A.13) and (A.14) we obtain that the following bound holds with $I\!P$-probability at least $1 - \mathrm{e}^{-\mathtt{x}}$:

$$I\!P^{\circ} \left(\sup_{\boldsymbol{\theta}, \boldsymbol{\theta}_1 \in \Theta_0(\mathtt{r})} |\mathcal{A}^{\circ}(\boldsymbol{\theta}, \boldsymbol{\theta}_1)| \le 4\mathtt{r} \left\{ \Delta_{\mathrm{W}}(\mathtt{r}, \mathtt{x})/3 + 6 \nu_0 \mathtt{r} \omega_1(\mathtt{r}) \, \mathfrak{Z}(\mathtt{x}) \right\} \right) \ge 1 - \mathrm{e}^{-\mathtt{x}} \tag{A.15}$$

for $\mathtt{r} \le \mathtt{r}_0$.

Theorems A.5 and A.1 say that the maximum likelihood estimators $\widetilde{\boldsymbol{\theta}}^{\circ}$ and $\widetilde{\boldsymbol{\theta}}$ get into the local vicinity $\Theta_0(\mathtt{r}_0)$ with exponentially high $I\!P^{\circ}$- and $I\!P$-probabilities

correspondingly. Therefore, taking $\boldsymbol{\theta} = \widetilde{\boldsymbol{\theta}}^{\circ}$ and $\boldsymbol{\theta}_1 = \widetilde{\boldsymbol{\theta}}$ in the last bound, we obtain with dominating probability:

$$\begin{aligned}
&\left| L^{\circ}(\widetilde{\boldsymbol{\theta}}^{\circ}) - L^{\circ}(\widetilde{\boldsymbol{\theta}}) - (\widetilde{\boldsymbol{\theta}}^{\circ} - \widetilde{\boldsymbol{\theta}})^{\top} \nabla_{\boldsymbol{\theta}} L^{\circ}(\widetilde{\boldsymbol{\theta}}) + \frac{1}{2} \| D_0 (\widetilde{\boldsymbol{\theta}}^{\circ} - \widetilde{\boldsymbol{\theta}}) \|^2 \right| \\
&\qquad \leq 4\mathtt{r} \left\{ \Delta_{\mathrm{W}}(\mathtt{r}_0, \mathtt{x})/3 + 6\nu_0 \mathtt{r}_0 \omega_1(\mathtt{r})\, \mathfrak{Z}(\mathtt{x}) \right\}.
\end{aligned}$$

Similarly bounds (A.13) and (A.14) imply:

$$\begin{aligned}
&\frac{1}{2} \left| \| \boldsymbol{\xi}^{\circ}(\widetilde{\boldsymbol{\theta}}) \|^2 - 2(\widetilde{\boldsymbol{\theta}}^{\circ} - \widetilde{\boldsymbol{\theta}})^{\top} \nabla_{\boldsymbol{\theta}} L^{\circ}(\widetilde{\boldsymbol{\theta}}) + \| D_0 (\widetilde{\boldsymbol{\theta}}^{\circ} - \widetilde{\boldsymbol{\theta}}) \|^2 \right| \\
&\qquad = \frac{1}{2} \left\| D_0^{-1} \nabla_{\boldsymbol{\theta}} L^{\circ}(\widetilde{\boldsymbol{\theta}}) - D_0 (\widetilde{\boldsymbol{\theta}}^{\circ} - \widetilde{\boldsymbol{\theta}}) \right\|^2 \\
&\qquad \leq 2 \left\{ \Delta_{\mathrm{W}}(\mathtt{r}_0, \mathtt{x})/3 + 6\nu_0 \mathtt{r}_0 \omega_1(\mathtt{r})\, \mathfrak{Z}(\mathtt{x}) \right\}^2. \qquad \text{(A.16)}
\end{aligned}$$

Therefore it holds with $I\!P$-probablity at least $1 - 4\mathrm{e}^{-\mathtt{x}}$:

$$\begin{aligned}
&I\!P^{\circ} \left(\left| L^{\circ}(\widetilde{\boldsymbol{\theta}}^{\circ}) - L^{\circ}(\widetilde{\boldsymbol{\theta}}) - \frac{1}{2} \| \boldsymbol{\xi}^{\circ}(\widetilde{\boldsymbol{\theta}}) \|^2 \right| \leq \Delta^{\circ}_{\mathrm{W}^2}(\mathtt{r}_0, \mathtt{x}) \right) \geq 1 - 4\mathrm{e}^{-\mathtt{x}}, \\
&\Delta^{\circ}_{\mathrm{W}^2}(\mathtt{r}_0, \mathtt{x}) \stackrel{\mathrm{def}}{=} 4\mathtt{r} \left\{ \Delta_{\mathrm{W}}(\mathtt{r}_0, \mathtt{x})/3 + 6\nu_0 \mathtt{r}_0 \omega_1(\mathtt{r})\, \mathfrak{Z}(\mathtt{x}) \right\} \\
&\qquad\qquad + 2 \left\{ \Delta_{\mathrm{W}}(\mathtt{r}_0, \mathtt{x})/3 + 6\nu_0 \mathtt{r}_0 \omega_1(\mathtt{r})\, \mathfrak{Z}(\mathtt{x}) \right\}^2.
\end{aligned}$$

For the second bound of the statement we use the similar approach as in Theorem 2.3 in Spokoiny (2013).

$$\begin{aligned}
&\left| \sqrt{2 \left\{ L^{\circ}(\widetilde{\boldsymbol{\theta}}^{\circ}) - L^{\circ}(\widetilde{\boldsymbol{\theta}}) \right\}} - \| D_0 (\widetilde{\boldsymbol{\theta}}^{\circ} - \widetilde{\boldsymbol{\theta}}) \| \right| \\
&\qquad \leq \frac{\left| 2 \left\{ L^{\circ}(\widetilde{\boldsymbol{\theta}}^{\circ}) - L^{\circ}(\widetilde{\boldsymbol{\theta}}) \right\} - \| D_0 (\widetilde{\boldsymbol{\theta}}^{\circ} - \widetilde{\boldsymbol{\theta}}) \|^2 \right|}{\| D_0 (\widetilde{\boldsymbol{\theta}}^{\circ} - \widetilde{\boldsymbol{\theta}}) \|} \\
&\qquad = \frac{\left| 2\mathcal{A}^{\circ}(\widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\theta}}^{\circ}) \right|}{\| D_0 (\widetilde{\boldsymbol{\theta}}^{\circ} - \widetilde{\boldsymbol{\theta}}) \|} \leq \sup_{\boldsymbol{\theta}, \boldsymbol{\theta}_1 \in \Theta_0(\mathtt{r}_0)} \frac{|2\mathcal{A}^{\circ}(\boldsymbol{\theta}, \boldsymbol{\theta}_1)|}{\| D_0 (\boldsymbol{\theta} - \boldsymbol{\theta}_1) \|} \\
&\qquad \overset{\text{by (A.11)}}{\leq} \sup_{\boldsymbol{\theta}, \boldsymbol{\theta}_1 \in \Theta_0(\mathtt{r}_0)} 2 \left\| D_0^{-1} \nabla_{\boldsymbol{\theta}} \mathcal{A}^{\circ}(\boldsymbol{\theta}, \boldsymbol{\theta}_1) \right\| \qquad \text{(A.17)} \\
&\qquad \overset{\text{by (A.13), (A.14)}}{\leq} 4\Delta_{\mathrm{W}}(\mathtt{r}_0, \mathtt{x})/3 + 24\nu_0 \mathtt{r}_0 \omega_1(\mathtt{r})\, \mathfrak{Z}(\mathtt{x}).
\end{aligned}$$

This together with (A.16) imply the final statement. □

Lemma A.1 (Check of the bootstrap equivalent of $(\boldsymbol{ED_2})$). *Conditions* $(\boldsymbol{Eb})$, $(\boldsymbol{\mathcal{L}_{0m}})$ *and* $(\boldsymbol{ED_{2m}})$ *imply for each* $\mathtt{r} > 0$, $\boldsymbol{\theta} \in \Theta_0(\mathtt{r})$, $\|\boldsymbol{\gamma}_j\| = 1$, $j = 1, 2$ *and all* $|\lambda| \leq \mathtt{g}_2(\mathtt{r})$ *with probability* $\geq 1 - \mathrm{e}^{-\mathtt{x}}$*:*

$$\sup_{\substack{\boldsymbol{\gamma}_j \in I\!R^p \\ \|\boldsymbol{\gamma}_j\| = 1}} \sum_{i=1}^{n} \log I\!E^{\circ} \exp \left\{ \frac{\lambda}{\omega_1(\mathtt{r})} \boldsymbol{\gamma}_1^{\top} D_0^{-1} \nabla^2_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}) D_0^{-1} \boldsymbol{\gamma}_2 (u_i - 1) \right\} \leq \frac{\lambda^2 \nu_0^2}{2}.$$

where

$$\omega_1(\mathtt{r}) = \omega_1 \stackrel{\text{def}}{=} \frac{\mathtt{C}_m(\mathtt{r})}{\sqrt{n}} + 2\omega\nu_0\sqrt{2\mathtt{x}} \tag{A.18}$$

In the case A.3.1 it holds for $\mathtt{r} \le \mathtt{r}_0$ $\omega_1(\mathtt{r}) = \mathtt{C}\mathtt{r}/n + \mathtt{C}\sqrt{\mathtt{x}/n}$.

Proof of Lemma A.1. Introduce the independent random scalar values for $i = 1, \ldots, n$ and $j = 1, 2$:

$$\mu_i(\boldsymbol{\theta}, \boldsymbol{\gamma}_j) \stackrel{\text{def}}{=} \boldsymbol{\gamma}_1^\top D_0^{-1} \nabla^2_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}) D_0^{-1} \boldsymbol{\gamma}_2.$$

It holds

$$\begin{aligned}
&\sum_{i=1}^n \log I\!\!E^\bullet \exp\left\{ \frac{\lambda}{\omega_1} \boldsymbol{\gamma}_1^\top D_0^{-1} \nabla^2_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}) D_0^{-1} \boldsymbol{\gamma}_2 (u_i - 1) \right\} \\
&\quad = \sum_{i=1}^n \log I\!\!E^\bullet \exp\left\{ \frac{\lambda}{\omega_1} \mu_i(\boldsymbol{\theta}, \boldsymbol{\gamma}_j)(u_i - 1) \right\} \\
&\quad \le \frac{\lambda^2 \nu_0^2}{2\omega_1^2} \sum_{i=1}^n \mu_i^2(\boldsymbol{\theta}, \boldsymbol{\gamma}_j),
\end{aligned} \tag{A.19}$$

here the inequality (A.19) follows from condition $\boldsymbol{(Eb)}$ if $\bigl|\lambda\mu_i(\boldsymbol{\theta}, \boldsymbol{\gamma}_j)\bigr| \le \mathtt{g}\omega_1$ for all $i = 1, \ldots, n$, which is true due to the arguments below. Let us consider $\mu_i(\boldsymbol{\theta}, \boldsymbol{\gamma}_j)$, for each $\boldsymbol{\theta} \in \Theta_0(\mathtt{r})$, $i = 1, \ldots n$ it holds:

$$\begin{aligned}
\bigl|\mu_i(\boldsymbol{\theta}, \boldsymbol{\gamma}_j)\bigr| &\le \| D_0^{-1} \nabla^2_{\boldsymbol{\theta}} I\!\!E \ell_i(\boldsymbol{\theta}) D_0^{-1} \| \\
&\quad + \| D_0^{-1} \left\{ \nabla^2_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}) - \nabla^2_{\boldsymbol{\theta}} I\!\!E \ell_i(\boldsymbol{\theta}) \right\} D_0^{-1} \|.
\end{aligned} \tag{A.20}$$

Condition $\boldsymbol{(ED_{2m})}$, which is a stronger version of $\boldsymbol{(ED_2)}$, implies that for all $i = 1, \ldots, n$, $\boldsymbol{\theta} \in \Theta_0(\mathtt{r})$ and each $\mathtt{r} > 0$ it holds with $I\!P$ -probability $\ge 1 - \mathrm{e}^{-\mathtt{x}}$

$$\| D_0^{-1} \left\{ \nabla^2_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}) - \nabla^2_{\boldsymbol{\theta}} I\!\!E \ell_i(\boldsymbol{\theta}) \right\} D_0^{-1} \| \le 2\omega\nu_0 \left(\frac{2\mathtt{x}}{n} \right)^{1/2}. \tag{A.21}$$

Indeed, by the exponential Chebyshev inequality for $\lambda > 0$

$$\begin{aligned}
&I\!\!P\left(\omega^{-1} \| D_0^{-1} \left\{ \nabla^2_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}) - \nabla^2_{\boldsymbol{\theta}} I\!\!E \ell_i(\boldsymbol{\theta}) \right\} D_0^{-1} \| \ge t \right) \\
&\quad \le I\!\!E \exp\left[-\lambda t + \omega^{-1} \lambda \| D_0^{-1} \left\{ \nabla^2_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}) - \nabla^2_{\boldsymbol{\theta}} I\!\!E \ell_i(\boldsymbol{\theta}) \right\} D_0^{-1} \| \right] \\
&\quad \stackrel{\text{by } \boldsymbol{(ED_{2m})}}{\le} \exp\left\{ -\lambda t + \lambda^2 \nu_0^2 / (2n) \right\}, \quad 0 < \lambda < \mathtt{g}_2(\mathtt{r}) \\
&\quad \le \exp\{-\mathtt{x}\},
\end{aligned}$$

here the last inequality holds under the assumption, that $\mathtt{g}_2(\mathtt{r})$ is large enough. In the case A.3.1 it holds $\mathtt{g}_2(\mathtt{r}) = \mathtt{C}n^{1/2}$, $\omega = \mathtt{C}n^{-1/2}$ and $\mathtt{x} = \mathtt{C}\log(n)$; $t^2 := 8\nu_0^2\mathtt{x}/n$

implies $\lambda t - \lambda^2 \nu_0^2/(2n) - \mathtt{x} \geq 0$ for $0 < \lambda < \mathtt{g}_2(\mathtt{r})$. For the deterministic term in (A.20) condition $(\boldsymbol{\mathcal{L}_{0m}})$ reads as:

$$\| D_0^{-1} \nabla^2_{\boldsymbol{\theta}} I\!\!E \ell_i(\boldsymbol{\theta}) D_0^{-1} \| \leq \frac{\mathtt{C}_m(\mathtt{r})}{n}. \tag{A.22}$$

Collecting the inequalities (A.19), (A.20), (A.21) and (A.22), we obtain:

$$\begin{aligned} \sum_{i=1}^n \log I\!\!E^\circ \exp\left\{ \frac{\lambda}{\omega_1} \boldsymbol{\gamma}_1^\top D_0^{-1} \nabla^2_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}) D_0^{-1} \boldsymbol{\gamma}_2 (u_i - 1) \right\} \\ \leq \frac{\lambda^2 \nu_0^2}{2} \frac{1}{{\omega_1}^2} \left\{ \frac{\mathtt{C}_m(\mathtt{r})}{\sqrt{n}} + 2\omega\nu_0 \sqrt{2\mathtt{x}} \right\}^2 \end{aligned}$$

Taking $\omega_1 = \omega_1(\mathtt{r})$ as in (A.18) implies the necessary statement. □

Theorem A.5 (Concentration of bootstrap MLE)**.** *Let the conditions of Theorems A.1 and A.6,* $(\boldsymbol{\mathcal{L}_{0m}})$ *and* $(\boldsymbol{ED_{2m}})$ *be fulfilled. If the following holds for* $\omega_1(\mathtt{r})$ *defined in* (A.18) *and the* $I\!\!P$*-random matrix* $\mathcal{B}^2 \stackrel{\text{def}}{=} D_0^{-1} \operatorname{Var}^\circ \{\nabla_{\boldsymbol{\theta}} L^\circ(\boldsymbol{\theta}^*)\} D_0^{-1}$

$$\begin{aligned} \mathtt{b}(\mathtt{r})\mathtt{r} \geq 2 \left\{ \mathfrak{z}_{\mathtt{qf}}(\mathtt{x}, I\!\!B) + \mathfrak{z}_{\mathtt{qf}}(\mathtt{x}, \mathcal{B}) + 6\nu_0 \, \mathfrak{z}(\mathtt{x}) \omega_1(\mathtt{r}_0) \mathtt{r}_0 \right\} \\ + \, 12\nu_0(\omega + \omega_1(\mathtt{r})) \, \mathfrak{z}(\mathtt{x} + \log(2\mathtt{r}/\mathtt{r}_0)) \quad \textit{for} \quad \mathtt{r} > \mathtt{r}_0, \end{aligned} \tag{A.23}$$

then it holds with $I\!\!P$*-probability* $\geq 1 - 3\mathrm{e}^{-\mathtt{x}}$

$$I\!\!P^\circ \left(\widetilde{\boldsymbol{\theta}}^\circ \notin \Theta_0(\mathtt{r}_0) \right) \leq 3\mathrm{e}^{-\mathtt{x}}.$$

Proof of Theorem A.5. We use the idea by Spokoiny (2013): if $\sup_{\boldsymbol{\theta} \in \Theta \setminus \Theta_0(\mathtt{r}_0)} \{ L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}^*) \} < 0$, then $\widetilde{\boldsymbol{\theta}} \in \Theta_0(\mathtt{r}_0)$. We apply it here for the the bootstrap objects: $L^\circ(\boldsymbol{\theta}) - L^\circ(\widetilde{\boldsymbol{\theta}})$ and $\widetilde{\boldsymbol{\theta}}^\circ$. Denote the stochastic part of the bootstrap likelihood process as $\boldsymbol{\zeta}^\circ(\boldsymbol{\theta}) \stackrel{\text{def}}{=} L^\circ(\boldsymbol{\theta}) - I\!\!E^\circ L^\circ(\boldsymbol{\theta})$. It holds

$$\begin{aligned} L^\circ(\boldsymbol{\theta}) - L^\circ(\widetilde{\boldsymbol{\theta}}) &= \boldsymbol{\zeta}^\circ(\boldsymbol{\theta}) - \boldsymbol{\zeta}^\circ(\widetilde{\boldsymbol{\theta}}) + I\!\!E^\circ L^\circ(\boldsymbol{\theta}) - I\!\!E^\circ L^\circ(\widetilde{\boldsymbol{\theta}}) \\ &= \boldsymbol{\zeta}^\circ(\boldsymbol{\theta}) - \boldsymbol{\zeta}^\circ(\widetilde{\boldsymbol{\theta}}) + L(\boldsymbol{\theta}) - L(\widetilde{\boldsymbol{\theta}}) \\ &= \{ \boldsymbol{\zeta}^\circ(\boldsymbol{\theta}) - \boldsymbol{\zeta}^\circ(\widetilde{\boldsymbol{\theta}}) \} + \{ L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}^*) \} + \{ L(\boldsymbol{\theta}^*) - L(\widetilde{\boldsymbol{\theta}}) \}. \end{aligned}$$

Here the last summand $\{ L(\boldsymbol{\theta}^*) - L(\widetilde{\boldsymbol{\theta}}) \}$ is non-positive by definition (1.2) of $\widetilde{\boldsymbol{\theta}}$. The following bound follows from the proof of Theorem 2.1 in Spokoiny (2013):

$$I\!\!P \left(\sup_{\boldsymbol{\theta} \in \Theta \setminus \Theta_0(\mathtt{r}_0)} \{ L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}^*) \} < \varrho(\mathtt{r}, \mathtt{x})\mathtt{r} + \mathtt{r} \, \mathfrak{z}_{\mathtt{qf}}(\mathtt{x}, I\!\!B) - \mathtt{r}^2 \mathtt{b}(\mathtt{r})/2 \right) \geq 1 - 3\mathrm{e}^{-\mathtt{x}},$$

$$\varrho(\mathtt{r}, \mathtt{x}) \stackrel{\text{def}}{=} 6\nu_0 \, \mathfrak{z}(\mathtt{x} + \log(2\mathtt{r}/\mathtt{r}_0))\omega.$$

Due to Lemma A.1 the process $\boldsymbol{\zeta}^{\circ}(\boldsymbol{\theta}) - \boldsymbol{\zeta}^{\circ}(\widetilde{\boldsymbol{\theta}})$ satisfies the conditions of Theorem A.1 in Spokoiny (2013), and it holds for $\mathtt{r} \geq \mathtt{r}_0$

$$\mathbb{P}^{\circ}\left(\sup_{\boldsymbol{\theta} \in \Theta_0(\mathtt{r})} \left| \boldsymbol{\zeta}^{\circ}(\boldsymbol{\theta}) - \boldsymbol{\zeta}^{\circ}(\widetilde{\boldsymbol{\theta}}) - (\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}})^{\top} \nabla_{\boldsymbol{\theta}} \boldsymbol{\zeta}^{\circ}(\widetilde{\boldsymbol{\theta}}) \right| \leq \varrho_1(\mathtt{r}, \mathtt{x}) \mathtt{r} \right) \geq 1 - \mathrm{e}^{-\mathtt{x}},$$

$$\varrho_1(\mathtt{r}, \mathtt{x}) \stackrel{\text{def}}{=} 6\nu_0 \, \mathfrak{z}(\mathtt{x} + \log(2\mathtt{r}/\mathtt{r}_0)) \omega_1(\mathtt{r}).$$

By Lemma A.2 and Theorem A.6 it holds with dominating probability

$$\begin{aligned} \sup_{\boldsymbol{\theta} \in \Theta_0(\mathtt{r})} \left| (\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}})^{\top} \nabla_{\boldsymbol{\theta}} \boldsymbol{\zeta}^{\circ}(\widetilde{\boldsymbol{\theta}}) \right| &\leq \mathtt{r} \| \boldsymbol{\xi}^{\circ}(\widetilde{\boldsymbol{\theta}}) \| \\ &\leq \mathtt{r} \left\{ \| \boldsymbol{\xi}^{\circ}(\boldsymbol{\theta}^*) \| + \| \boldsymbol{\xi}^{\circ}(\widetilde{\boldsymbol{\theta}}) - \boldsymbol{\xi}^{\circ}(\boldsymbol{\theta}^*) \| \right\} \\ &\leq \mathtt{r} \left\{ \mathfrak{z}_{\mathtt{qf}}(\mathtt{x}, \mathcal{B}) + 6\nu_0 \, \mathfrak{z}(\mathtt{x}) \omega_1(\mathtt{r}_0) \mathtt{r}_0 \right\}. \end{aligned}$$

Finally we have:

$$\begin{aligned} &\sup_{\boldsymbol{\theta} \in \Theta \setminus \Theta_0(\mathtt{r}_0)} \left\{ L^{\circ}(\boldsymbol{\theta}) - L^{\circ}(\widetilde{\boldsymbol{\theta}}) \right\} \\ &\leq \sup_{\boldsymbol{\theta} \in \Theta \setminus \Theta_0(\mathtt{r}_0)} \left\{ L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}^*) \right\} + \sup_{\substack{\boldsymbol{\theta} \in \Theta_0(\mathtt{r}), \\ \mathtt{r} \geq \mathtt{r}_0}} \left\{ \boldsymbol{\zeta}^{\circ}(\boldsymbol{\theta}) - \boldsymbol{\zeta}^{\circ}(\widetilde{\boldsymbol{\theta}}) \right\} \\ &\leq \mathtt{r} \, \mathfrak{z}_{\mathtt{qf}}(\mathtt{x}, \mathcal{B}) + \mathtt{r} \, \mathfrak{z}_{\mathtt{qf}}(\mathtt{x}, I\!\!B) + \varrho_1(\mathtt{r}, \mathtt{x}) \mathtt{r} + \varrho(\mathtt{r}, \mathtt{x}) \mathtt{r} - \mathtt{r}^2 \mathtt{b}(\mathtt{r})/2 + 6\nu_0 \, \mathfrak{z}(\mathtt{x}) \omega_1(\mathtt{r}_0) \mathtt{r} \mathtt{r}_0, \end{aligned}$$

which implies the condition (A.45) in the statement. □

Remark A.1. Condition (A.45) imposed for the bootstrap MLE concentration result is stronger than condition (A.2) for the concentration of $\boldsymbol{Y}$ - MLE, and (A.45) implies the latter one.

The following lemma had already been derived in the proof of Theorem A.4: see the bound (A.14). We formulate it separately, since it is used again in another statements.

Lemma A.2. *Let the conditions of Lemma A.1 be fulfilled, then it holds for* $\mathtt{r} \leq \mathtt{r}_0$ *with* $\mathbb{P}$ *-probability* $\geq 1 - \mathrm{e}^{-\mathtt{x}}$

$$\mathbb{P}^{\circ}\left(\sup_{\boldsymbol{\theta} \in \Theta_0(\mathtt{r})} \| \boldsymbol{\xi}^{\circ}(\boldsymbol{\theta}) - \boldsymbol{\xi}^{\circ}(\boldsymbol{\theta}^*) \| \leq \Delta_{\xi}^{\circ}(\mathtt{r}, \mathtt{x}) \right) \geq 1 - \mathrm{e}^{-\mathtt{x}},$$

where

$$\Delta_{\xi}^{\circ}(\mathtt{r}, \mathtt{x}) \stackrel{\text{def}}{=} 6\nu_0 \, \mathfrak{z}(\mathtt{x}) \omega_1(\mathtt{r}) \mathtt{r} \tag{A.24}$$

In the case A.3.1 it holds for the bounding term

$$\Delta_{\xi}^{\circ}(\mathtt{r}_0, \mathtt{x}) \leq \mathtt{C} \frac{p + \mathtt{x}}{\sqrt{n}} \sqrt{\mathtt{x}}.$$

Theorem A.6 (Deviation bound for the bootstrap quadratic form)**.** *Let conditions* $(\boldsymbol{Eb})$, $(\boldsymbol{\mathcal{I}})$, $(\boldsymbol{SD_1})$, $(\boldsymbol{\mathcal{I}_B})$ *be fulfilled, then for* $\mathtt{g} \geq \sqrt{2\,\mathrm{tr}(\mathcal{B}^2)}$ *it holds:*

$$I\!P^{\circ}\left(\|\boldsymbol{\xi}^{\circ}(\boldsymbol{\theta}^*)\|^2 \leq \mathfrak{Z}^2_{\mathtt{qf}}(\mathtt{x},\mathcal{B})\right) \geq 1 - 2\mathrm{e}^{-\mathtt{x}} - 8.4\mathrm{e}^{-\mathtt{x}_c(\mathcal{B})},$$

where

$$\mathcal{B}^2 \stackrel{\mathrm{def}}{=} D_0^{-1}\mathcal{V}^2(\boldsymbol{\theta}^*)D_0^{-1}, \quad \mathcal{V}^2(\boldsymbol{\theta}^*) \stackrel{\mathrm{def}}{=} \mathrm{Var}^{\circ}\,\nabla_{\boldsymbol{\theta}} L^{\circ}(\boldsymbol{\theta}^*), \tag{A.25}$$

$\mathfrak{Z}_{\mathtt{qf}}(\mathtt{x},\cdot)$ *and* $\mathtt{x}_c(\cdot)$ *are defined respectively in* (A.6) *and* (A.7)*. Similarly to* (A.8) *it holds for* $\mathtt{x} \leq \mathtt{x}_c(\mathcal{B})$*:*

$$\mathfrak{Z}^2_{\mathtt{qf}}(\mathtt{x},\mathcal{B}) \leq \mathfrak{a}^{\circ 2}(p + 6\mathtt{x}) \tag{A.26}$$

$$\text{for } \mathfrak{a}^{\circ 2} \stackrel{\mathrm{def}}{=} (1 + \delta^2_{\mathcal{V}}(\mathtt{x}))(\mathfrak{a}^2 + \mathfrak{a}^2_B), \tag{A.27}$$

and $\delta^2_{\mathcal{V}}(\mathtt{x})$ *given in* (D.36) *(see Section D.1.4 on Bernstein matrix inequality).*

Proof of Theorem A.6. This result is the bootstrap equivalent of Theorem A.3. For the $\boldsymbol{Y}$-world it demands condition $(\boldsymbol{ED_0})$ to be fulfilled. Let us check whether the bootstrap equivalent of $(\boldsymbol{ED_0})$ holds. It reads as follows: *there exist constants* $\mathtt{g}^{\circ} > 0$, $\nu_0^{\circ} \geq 1$ *such that for the positive-definite symmetric matrix* $\mathcal{V}^2(\boldsymbol{\theta}^*)$ *it holds for all* $|\lambda| \leq \mathtt{g}^{\circ}$

$$\sup_{\boldsymbol{\gamma} \in I\!R^p} \log I\!E^{\circ} \exp\left\{\lambda \frac{\boldsymbol{\gamma}^{\top}\{\nabla_{\boldsymbol{\theta}} L^{\circ}(\boldsymbol{\theta}^*) - \nabla_{\boldsymbol{\theta}} I\!E^{\circ} L^{\circ}(\boldsymbol{\theta}^*)\}}{\|\mathcal{V}(\boldsymbol{\theta}^*)\boldsymbol{\gamma}\|}\right\} \leq \nu_0^{\circ 2}\lambda^2/2.$$

By definition $\mathcal{V}^2(\boldsymbol{\theta}^*) = \sum_{i=1}^n \nabla_{\boldsymbol{\theta}}\ell_i(\boldsymbol{\theta}^*)\nabla_{\boldsymbol{\theta}}\ell_i(\boldsymbol{\theta}^*)^{\top}$. Let us introduce the independent $I\!P$-random variables $s_i(\boldsymbol{\gamma}) \stackrel{\mathrm{def}}{=} \boldsymbol{\gamma}^{\top}\nabla_{\boldsymbol{\theta}}\ell_i(\boldsymbol{\theta}^*)/\|\mathcal{V}(\boldsymbol{\theta}^*)\boldsymbol{\gamma}\|$ for $i = 1,\ldots,n$. It holds $\sum_{i=1}^n s_i^2(\boldsymbol{\gamma}) = 1$, hence $\max_{1\leq i\leq n}|s_i| \leq 1$. Condition $(\boldsymbol{Eb})$ implies:

$$\begin{aligned}
&\log I\!E^{\circ} \exp\left\{\lambda \frac{\boldsymbol{\gamma}^{\top}\{\nabla_{\boldsymbol{\theta}} L^{\circ}(\boldsymbol{\theta}^*) - \nabla_{\boldsymbol{\theta}} I\!E^{\circ} L^{\circ}(\boldsymbol{\theta}^*)\}}{\|\mathcal{V}(\boldsymbol{\theta}^*)\boldsymbol{\gamma}\|}\right\} \\
&\quad = \sum_{i=1}^n \log I\!E^{\circ} \exp\{\lambda s_i(\boldsymbol{\gamma})(u_i - 1)\} \\
&\quad \leq \frac{{\nu_0}^2\lambda^2}{2}\sum_{i=1}^n s_i^2(\boldsymbol{\gamma}) = {\nu_0}^2\lambda^2/2, \qquad |\lambda| \leq \mathtt{g}.
\end{aligned}$$

Thus the bootstrap equivalent for the condition $(\boldsymbol{ED_0})$ is fulfilled with the same constants $\nu_0, \mathtt{g}$, and the theorem's statements holds as well as for Theorem A.3.

The inequality (A.26) follows from conditions $(\boldsymbol{\mathcal{I}})$, $(\boldsymbol{\mathcal{I}_B})$, $(\boldsymbol{SD_1})$ and Bernstein matrix inequality by Tropp (2012) (see Section D.1.4):

$$\|D_0^{-1}\mathcal{V}_0^2(\boldsymbol{\theta}^*)D_0^{-1}\| \leq \|D_0^{-1}H_0\|^2(1 + \delta^2_{\mathcal{V}}(\mathtt{x})) \leq (1 + \delta^2_{\mathcal{V}}(\mathtt{x}))(\mathfrak{a}^2 + \mathfrak{a}^2_B).$$

□

A.3 Some frequently used models

Below we specify the results of Sections A.1, A.2 and conditions from Section 2.5 for some common models: the case of i.i.d. observations, generalised linear model, and linear quantile regression. We also show the dependence of the non-asymptotic bounds on n and p. Spokoiny (2012a) considered examples for the i.i.d. observations, generalised linear model and linear median regression, so this section has some overlapping with Section 5 in Spokoiny (2012a).

Throughout this section it is supposed that $\mathtt{x} \le \mathtt{C} \log n$.

A.3.1 I.i.d. observations (IID)

The observations $Y_1, \ldots, Y_n$ are independent and identically distributed.

Recall the notations for the marginal log-likelihood process and its stochastic part: $\ell_i(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \log\left\{\frac{d\mathbb{P}_{\boldsymbol{\theta}}}{d\mu_0}(Y_i)\right\}$, $\zeta_i(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \ell_i(\boldsymbol{\theta}) - \mathbb{E}\ell_i(\boldsymbol{\theta})$, $i = 1, \ldots, n$.

Lemma A.3. *Let for the IID case the conditions below be fulfilled:*

($\boldsymbol{ED_0}$ & IID) *There exist a positive-definite symmetric matrix v_0^2 and constants $\overline{\mathtt{g}} > 0, \nu_0 \ge 1$ such that* $\operatorname{Var}\{\nabla_{\boldsymbol{\theta}}\zeta_i(\boldsymbol{\theta}^*)\} \le v_0^2$ *and*

$$\sup_{\boldsymbol{\gamma} \in \mathbb{R}^p} \log \mathbb{E} \exp\left\{\lambda \frac{\boldsymbol{\gamma}^\top \nabla_{\boldsymbol{\theta}}\zeta_i(\boldsymbol{\theta}^*)}{\|v_0\boldsymbol{\gamma}\|}\right\} \le \nu_0^2\lambda^2/2, \qquad |\lambda| \le \overline{\mathtt{g}}.$$

($\boldsymbol{ED_2}$ & IID) *There exists a constant $\overline{\omega} > 0$ and for each $\mathtt{r} > 0$ a constant $\overline{\mathtt{g}}_2(\mathtt{r})$ such that it holds for all $\boldsymbol{\theta} \in \Theta_0(\mathtt{r})$, $j = 1, 2$, and $d_0^2 \stackrel{\text{def}}{=} -\nabla^2_{\boldsymbol{\theta}}\mathbb{E}\ell_i(\boldsymbol{\theta}^*)$*

$$\sup_{\substack{\boldsymbol{\gamma}_j \in \mathbb{R}^p \\ \|\boldsymbol{\gamma}_j\| \le 1}} \log \mathbb{E} \exp\left\{\overline{\omega}^{-1}\lambda \boldsymbol{\gamma}_1^\top d_0^{-1} \nabla^2_{\boldsymbol{\theta}}\zeta_i(\boldsymbol{\theta}) d_0^{-1}\boldsymbol{\gamma}_2\right\} \le \nu_0^2\lambda^2/2, \qquad |\lambda| \le \overline{\mathtt{g}}_2(\mathtt{r}),$$

($\boldsymbol{\mathcal{L}_{3m}}$ & IID) *For each $\mathtt{r} \ge 0$, and for all $\boldsymbol{\theta} \in \Theta_0(\mathtt{r})$ and $\boldsymbol{\gamma} \in \mathbb{R}^p : \|\boldsymbol{\gamma}\| = 1$ there exists a constant $\mathtt{C}_{3m} \ge 0$ such that*

$$\|D_0^{-1}\boldsymbol{\gamma}^\top \nabla^3_{\boldsymbol{\theta}}\mathbb{E}\ell_i(\boldsymbol{\theta}) D_0^{-1}\| \le \mathtt{C}_{3m} n^{-1}.$$

($\boldsymbol{\mathcal{L}}\mathtt{rc}$) *For each $\mathtt{r} \ge \mathtt{r}_0$ and $\forall \boldsymbol{\theta} \in \Theta_0(\mathtt{r})$* it holds for some value $\mathtt{C}_\mathtt{b}(\mathtt{r}) > 0$ s.t. $\mathtt{r}\mathtt{C}_\mathtt{b}(\mathtt{r}) \to +\infty$ with $\mathtt{r} \to +\infty$ and

$$\left\|D_0^{-1} D^2(\boldsymbol{\theta}) D_0^{-1}\right\| \ge \mathtt{C}_\mathtt{b}(\mathtt{r}).$$

Then the statements from Sections A.1 and A.2 and their conditions from Section 2.5 depend on the values from Table A.1.

Table A.1: The IID case

Source of the values	Corresponding values
$(\boldsymbol{ED_0})\,, (\boldsymbol{ED_2})$	$\mathtt{g} = \sqrt{n}\overline{\mathtt{g}}, \quad \mathtt{g}_2(\mathtt{r}) = \sqrt{n}\overline{\mathtt{g}}_2(\mathtt{r}), \quad \omega = \overline{\omega}/\sqrt{n},$
$(\mathcal{L}_0)$	$\delta(\mathtt{r}) = \mathtt{C}\mathtt{r}/\sqrt{n},$
$(\mathcal{L}\mathtt{r})$, Th. A.1, A.2, A.5	$\mathtt{b}(\mathtt{r}) = \mathtt{C}_\mathtt{b}(\mathtt{r}), \quad \mathfrak{Z}(\mathtt{x}) = \mathtt{C}\sqrt{p+\mathtt{x}}, \quad \mathtt{r}_0 \geq \mathtt{C}\sqrt{p+\mathtt{x}},$
Th. A.3	$\mathfrak{Z}^2_{\mathtt{qf}}(\mathtt{x}, I\!B) \leq \mathtt{C}\mathfrak{a}^2(p+6\mathtt{x})\,,$
Th. A.6	$\mathfrak{Z}^2_{\mathtt{qf}}(\mathtt{x}, \mathcal{B}) \leq \mathtt{C}\mathfrak{a}^{\circ 2}(p+6\mathtt{x})$ for $\mathfrak{a}^{\circ 2}$ from (A.27),
Th. A.2, A.4	$\Delta_{\mathrm{W}}(\mathtt{r}_0, \mathtt{x}), \Delta^\circ_{\mathrm{W}}(\mathtt{r}_0, \mathtt{x})\mathtt{x}^{-1/2} \leq \mathtt{C}\dfrac{p+\mathtt{x}}{\sqrt{n}},$
	$\Delta_{\mathrm{W}^2}(\mathtt{r}_0, \mathtt{x}), \Delta^\circ_{\mathrm{W}^2}(\mathtt{r}_0, \mathtt{x})\mathtt{x}^{-1/2} \leq \mathtt{C}\dfrac{(p+\mathtt{x})^{3/2}}{\sqrt{n}},$
Th. A.4, L. A.2	$\omega_1(\mathtt{r}, \mathtt{x}) \leq \mathtt{C}\dfrac{\mathtt{r}}{n} + \mathtt{C}\sqrt{\dfrac{\mathtt{x}}{n}}, \quad \Delta^\circ_\xi(\mathtt{r}_0, \mathtt{x}) \leq \mathtt{C}\dfrac{p+\mathtt{x}}{\sqrt{n}}\sqrt{\mathtt{x}},$
$(\mathbf{SmB})\,, (\mathcal{I}_{\boldsymbol{B}})\,, (\mathcal{L}_{0\boldsymbol{m}})$	$\delta^2_{\mathrm{smb}} = \mathfrak{a}^2_B = 0, \quad \mathtt{C}_m(\mathtt{r}) \leq \mathtt{C}\dfrac{\mathtt{r}}{\sqrt{n}} + \mathtt{C}.$

Proof of Lemma A.3. Take $V_0^2 = \mathrm{Var}\left\{\nabla_{\boldsymbol{\theta}}\zeta(\boldsymbol{\theta}^*)\right\}$ and $v_0^2 \stackrel{\mathrm{def}}{=} \mathrm{Var}\left\{\nabla_{\boldsymbol{\theta}}\zeta_1(\boldsymbol{\theta}^*)\right\}$, then $V_0^2 = nv_0^2$ due to the i.i.d. property of Y_i. Similarly $D_0^2 = nd_0^2$, $D^2(\boldsymbol{\theta}) = nd^2(\boldsymbol{\theta})$ for $d_0^2 \stackrel{\mathrm{def}}{=} -\nabla^2_{\boldsymbol{\theta}} I\!E\ell_1(\boldsymbol{\theta}^*)$, $D^2(\boldsymbol{\theta}) \stackrel{\mathrm{def}}{=} -\nabla^2_{\boldsymbol{\theta}} I\!EL(\boldsymbol{\theta})$, $d^2(\boldsymbol{\theta}) \stackrel{\mathrm{def}}{=} -\nabla^2_{\boldsymbol{\theta}} I\!E\ell_1(\boldsymbol{\theta})$. For the condition $(\boldsymbol{ED_0})$ it holds by independence of the observations and by $(\boldsymbol{ED_0}\,\&\,\mathbf{IID})$:

$$\begin{aligned}\log I\!E \exp\left\{\lambda \frac{\boldsymbol{\gamma}^\top \nabla_{\boldsymbol{\theta}}\zeta(\boldsymbol{\theta}^*)}{\|V_0\boldsymbol{\gamma}\|}\right\} &= \sum\nolimits_{i=1}^n \log I\!E \exp\left\{\lambda \frac{\boldsymbol{\gamma}^\top \nabla_{\boldsymbol{\theta}}\zeta_i(\boldsymbol{\theta}^*)}{\sqrt{n}\|v_0\boldsymbol{\gamma}\|}\right\} \\ &\leq n\frac{\lambda^2\nu_0^2}{2n} \quad \text{for } |\lambda| \leq \sqrt{n}\overline{\mathtt{g}}.\end{aligned}$$

Similarly for the condition $(\boldsymbol{ED_2})$:

$$\begin{aligned}
&\log I\!\!E \exp\left\{\frac{\lambda}{\omega}\boldsymbol{\gamma}_1^\top D_0^{-1}\nabla^2_{\boldsymbol{\theta}}\zeta(\boldsymbol{\theta})D_0^{-1}\boldsymbol{\gamma}_2\right\} \\
&\quad = \sum\nolimits_{i=1}^n \log I\!\!E \exp\left\{\frac{\lambda}{\sqrt{n}}\frac{1}{\omega\sqrt{n}}\boldsymbol{\gamma}_1^\top d_0^{-1}\nabla^2_{\boldsymbol{\theta}}\zeta_i(\boldsymbol{\theta})d_0^{-1}\boldsymbol{\gamma}_2\right\} \\
&\quad \le n\frac{\lambda^2\nu_0^2}{2n} \quad \text{for } |\lambda| \le \sqrt{n}\overline{\mathtt{g}}_2(\mathtt{r}),\ \omega\sqrt{n} = \overline{\omega}.
\end{aligned} \tag{A.28}$$

In the condition $(\boldsymbol{\mathcal{L}_0})$ it holds for $\mathtt{r} \le \mathtt{r}_0$, $\boldsymbol{\theta} \in \Theta_0(\mathtt{r})$ and some $\overline{\boldsymbol{\theta}} \in \Theta_0(\mathtt{r})$

$$\begin{aligned}
&\|D_0^{-1}D^2(\boldsymbol{\theta})D_0^{-1} - \boldsymbol{I}_p\| = \|D_0^{-1}(\boldsymbol{\theta}^* - \boldsymbol{\theta})^\top\nabla^3_{\boldsymbol{\theta}} I\!\!E L(\overline{\boldsymbol{\theta}})D_0^{-1}\| \\
&= \|D_0^{-1}(\boldsymbol{\theta}^* - \boldsymbol{\theta})^\top D_0 D_0^{-1}\nabla^3_{\boldsymbol{\theta}} I\!\!E L(\overline{\boldsymbol{\theta}})D_0^{-1}\| \\
&\le \mathtt{r}\|D_0^{-1}\|\|D_0^{-1}\mathbf{1}_p^\top\nabla^3_{\boldsymbol{\theta}} I\!\!E L(\overline{\boldsymbol{\theta}})D_0^{-1}\| \le \mathtt{C}\mathtt{r}/\sqrt{n} \quad \text{(by condition } (\boldsymbol{\mathcal{L}_{3m}}\,\&\,\mathbf{IID})\text{)},
\end{aligned} \tag{A.29}$$

therefore $\delta(\mathtt{r}) = \mathtt{C}\mathtt{r}/\sqrt{n}$.

The value $\mathtt{b}(\mathtt{r})$ from condition $(\boldsymbol{\mathcal{L}\mathtt{r}})$ can be taken equal to $\mathtt{C}_\mathtt{b}$ due to condition $(\mathcal{L}\mathtt{rc})$. By definition (A.5) $\mathfrak{Z}(\mathtt{x}) = \mathtt{C}\sqrt{p+\mathtt{x}}$. The value $\mathtt{x} = \mathtt{C}\log n$ corresponds to the first two regimes in (A.6). By condition $(\boldsymbol{\mathcal{I}})$ $\operatorname{tr}(I\!\!B^2) \le \mathfrak{a}^2 p$, $\operatorname{tr}(I\!\!B^4) \le \mathfrak{a}^4 p$ and $\lambda(I\!\!B) \le \mathfrak{a}^2$, hence $\mathfrak{Z}^2_{\mathtt{qf}}(\mathtt{x}, I\!\!B) \le \mathtt{C}\mathfrak{a}^2(p+6\mathtt{x})$. Substitution of the obtained values in the condition (A.2) of Theorem A.1 on concentration of the MLE $\widetilde{\boldsymbol{\theta}}$ yields $\mathtt{r}_0 \ge \mathtt{C}\sqrt{p+\mathtt{x}}$. Similarly for the error terms of the Wilks approximations given in (A.3) and (A.4) it holds for $\mathtt{r} \le \mathtt{r}_0$:

$$\Delta_{\mathrm{W}}(\mathtt{r},\mathtt{x}) \le \mathtt{C}\,\frac{p+\mathtt{x}}{\sqrt{n}}, \quad \Delta_{\mathrm{W}^2}(\mathtt{r},\mathtt{x}) \le \mathtt{C}\sqrt{\frac{(p+\mathtt{x})^3}{n}}.$$

Now let us consider the statements from the bootstrap part. Similarly to (A.28) condition $(\boldsymbol{ED_2}\,\&\,\mathbf{IID})$ implies that $(\boldsymbol{ED_{2m}})$ is fulfilled with the same values ω and $\mathtt{g}$ as in $(\boldsymbol{ED_2})$.

$\mathtt{C}_m(\mathtt{r})$ from condition $(\boldsymbol{\mathcal{L}_{0m}})$ is bounded with $\mathtt{C}(\mathtt{r}/\sqrt{n}+1)$. Indeed, for $\boldsymbol{\theta} \in \Theta_0(\mathtt{r})$ and some $\overline{\boldsymbol{\theta}} \in \Theta_0(\mathtt{r})$ it holds due to condition $(\boldsymbol{\mathcal{L}_{3m}}\,\&\,\mathbf{IID})$

$$\begin{aligned}
&\left\|D_0^{-1}\nabla^2_{\boldsymbol{\theta}} I\!\!E\ell_i(\boldsymbol{\theta})D_0^{-1}\right\| \\
&\quad \le \|D_0^{-1}\nabla^2_{\boldsymbol{\theta}} I\!\!E\ell_i(\boldsymbol{\theta}^*)D_0^{-1}\| + \|D_0^{-1}(\boldsymbol{\theta}^* - \boldsymbol{\theta})^\top D_0 D_0^{-1}\nabla^3_{\boldsymbol{\theta}} I\!\!E\ell_i(\overline{\boldsymbol{\theta}})D_0^{-1}\| \\
&\quad \le n^{-1} + \mathtt{C}\mathtt{r}n^{-3/2}.
\end{aligned} \tag{A.30}$$

This implies by definitions (A.18) and (A.10) for $\mathtt{r} \le \mathtt{r}_0$: $\omega_1(\mathtt{r},\mathtt{x}) = \mathtt{C}\mathtt{r}/n + \mathtt{C}\sqrt{\mathtt{x}/n}$ and

$$\Delta^\circ_{\mathrm{W}}(\mathtt{r},\mathtt{x}) \le \mathtt{C}\,\frac{p+\mathtt{x}}{\sqrt{n}}\sqrt{\mathtt{x}}, \quad \Delta^\circ_{\mathrm{W}^2}(\mathtt{r},\mathtt{x}) \le \mathtt{C}\sqrt{\frac{(p+\mathtt{x})^3}{n}}\sqrt{\mathtt{x}}.$$

The relations above and definition (A.24) imply $\Delta_{\xi}^{\circ}(\mathtt{r},\mathtt{x}) \leq \mathtt{C}\,(p+\mathtt{x})\,\mathtt{x}^{1/2} n^{-1/2}$ for $\mathtt{r} \leq \mathtt{r}_0$. Similarly to $\mathfrak{Z}^2_{\mathtt{qf}}(\mathtt{x}, I\!B_k)$ it holds $\mathfrak{Z}^2_{\mathtt{qf}}(\mathtt{x}, \mathcal{B}) \leq \mathtt{C}\mathfrak{a}^{\circ 2}\,(p+6\mathtt{x})$ for $\mathfrak{a}^{\circ 2}$ given in (A.27). By definition (1.3) of $\boldsymbol{\theta}^*$ $\nabla_{\boldsymbol{\theta}} I\!E L(\boldsymbol{\theta}^*) = n \nabla_{\boldsymbol{\theta}} I\!E \ell_1(\boldsymbol{\theta}^*) = 0$, therefore $B_0^2 = 0$ (see def. (1.10)), and it can be taken $\delta^2_{\mathrm{smb}} = \mathfrak{a}_B^2 = 0$. Condition $(\boldsymbol{SD_1})$ reads as

$$\left\| v_0^{-1} \left\{ \nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}^*) \nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}^*)^\top \right\} v_0^{-1} - \boldsymbol{I}_p \right\| \leq \delta_v^2 n$$

with dominating probability. □

A.3.2 Generalized Linear Model (GLM)

Here we consider the Generalized Linear Model, introduced by Nelder and Wedderburn (1972). Let the parametric probability distribution family $\{I\!P_\upsilon\}$ be an exponential family with a canonical parametrisation. The log-density for this family can be expressed as

$$\ell(\upsilon) = y\upsilon - h(\upsilon)$$

for a convex function $h(\cdot)$. Table A.2 shows some particular examples of $\{I\!P_\upsilon\}$ and $h(\cdot)$. Taking $\{I\!P_\upsilon\}$ as a parametric family and $\Psi_i^\top \boldsymbol{\theta}$ as linear predictors for some

Table A.2: Examples of the GLM

$I\!P_\upsilon$	$h(\upsilon)$	$h'(\upsilon)$ (natural parameter)	$h''(\upsilon)$
$\mathcal{N}(\upsilon, 1)$	$\upsilon^2/2$	υ	1
$Exp(-\upsilon)$	$-\log(-\upsilon)$	$-1/\upsilon$	$1/\upsilon^2$
$Pois(\mathrm{e}^\upsilon)$	e^υ	e^υ	e^υ
$Binom\left(1, \frac{\mathrm{e}^\upsilon}{\mathrm{e}^\upsilon+1}\right)$	$\log(\mathrm{e}^\upsilon + 1)$	$\frac{\mathrm{e}^\upsilon}{\mathrm{e}^\upsilon+1}$	$\frac{\mathrm{e}^\upsilon}{(\mathrm{e}^\upsilon+1)^2}$

deterministic regressors $\Psi_i \in I\!R^p$ yields the following quasi log-likehood function:

$$L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \left\{ Y_i \Psi_i^\top \boldsymbol{\theta} - h(\Psi_i^\top \boldsymbol{\theta}) \right\}. \tag{A.31}$$

Let us recall that the true distribution $I\!P$ of the data sample $\boldsymbol{Y}$ is not required to belong to $\{I\!P_\upsilon\}$, and the true parameter $\boldsymbol{\theta}^*$ is defined by projection as in (1.3).

Lemma A.4. *Consider the Generalized Linear Model s.t. the parameter's domain Θ is compact and $h(\Psi_i^\top \boldsymbol{\theta})$ is three times differentiable on it. If the conditions below are fulfilled:*

(ED_0 & GLM) *There exist positive constants* $\sigma_1^2, \ldots, \sigma_n^2$ *and* $\overline{\mathtt{g}} > 0, \nu_0 \geq 1$ *such that for each* $i = 1, \ldots, n$ $\operatorname{Var} Y_i \leq \sigma_i^2$ *and*

$$\log I\!\!E \exp\left\{\lambda \frac{Y_i - I\!\!E Y_i}{\sigma_i}\right\} \leq \nu_0^2 \lambda^2/2, \qquad |\lambda| \leq \overline{\mathtt{g}}.$$

($\mathbf{h_0}$ & GLM) *For each* $\mathtt{r} \geq 0$ *there exists a value* $\mathtt{C}_{h_0}(\mathtt{r}) \geq 0$ *such that* $\forall\, \boldsymbol{\theta} \in \Theta_0(\mathtt{r})$ *and* $\Psi_i^\top \boldsymbol{\theta}$, $i = 1, \ldots, n$ *in the domain of the function* $h(\cdot)$ *it holds:*

$$\max_{\substack{1 \leq i \leq n, \\ h''(\Psi_i^\top \boldsymbol{\theta}^*) > 0}} \frac{\bigl|h'''(\Psi_i^\top \boldsymbol{\theta})\bigr|}{\bigl|h''(\Psi_i^\top \boldsymbol{\theta}^*)\bigr|^2} \leq \mathtt{C}_{h_0}(\mathtt{r}),$$

and $\mathtt{C}_{h_0}(\mathtt{r}) \leq \mathtt{C}_{h_0} = const$ *for* $\mathtt{r} \leq \mathtt{r}_0$.

($\mathbf{h_r}$ & GLM) *For each* $\mathtt{r} > \mathtt{r}_0$ *there exists a value* $\mathtt{C}_h(\mathtt{r}) > 0$ *s.t.* $\mathtt{r}\mathtt{C}_h(\mathtt{r}) \to +\infty$ *with* $\mathtt{r} \to +\infty$, *and* $\forall\, \boldsymbol{\theta} \in \Theta_0(\mathtt{r})$ *it holds*

$$\min_{\substack{1 \leq i \leq n, \\ h''(\Psi_i^\top \boldsymbol{\theta}^*) > 0}} \frac{\bigl|h''(\Psi_i^\top \boldsymbol{\theta})\bigr|}{\bigl|h''(\Psi_i^\top \boldsymbol{\theta}^*)\bigr|} \geq \mathtt{C}_h(\mathtt{r}).$$

Then the statements from Sections A.1 and A.2, and their conditions from Section 2.5 depend on the values from Table A.3 with

$$\frac{1}{\sqrt{N_\sigma}} \stackrel{\text{def}}{=} \max_{1 \leq i \leq n} \bigl\|\sigma_i V_0^{-1} \Psi_i\bigr\| \leq 1, \quad \frac{1}{\sqrt{N_h}} \stackrel{\text{def}}{=} \max_{1 \leq i \leq n} \bigl\|h''(\Psi_i^\top \boldsymbol{\theta}^*) D_0^{-1} \Psi_i\bigr\| \leq 1.$$

Proof of Lemma A.4. It holds:

$$\nabla_{\boldsymbol{\theta}} \zeta(\boldsymbol{\theta}) = \sum\nolimits_{i=1}^n \Psi_i \left(Y_i - I\!\!E Y_i\right), \tag{A.32}$$

$$V_0^2 \stackrel{\text{def}}{=} \operatorname{Var} \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}^*) = \sum\nolimits_{i=1}^n \Psi_i \Psi_i^\top \operatorname{Var} Y_i, \tag{A.33}$$

$$D^2(\boldsymbol{\theta}) \stackrel{\text{def}}{=} -\nabla^2_{\boldsymbol{\theta}} I\!\!E L(\boldsymbol{\theta}) = \sum\nolimits_{i=1}^n \Psi_i \Psi_i^\top h''(\Psi_i^\top \boldsymbol{\theta}). \tag{A.34}$$

Due to independence of the observations $Y_1, \ldots, Y_n$ and condition **(ED_0 & GLM)** it holds

$$\begin{aligned}
\log I\!\!E \exp\left\{\lambda \frac{\boldsymbol{\gamma}^\top \nabla_{\boldsymbol{\theta}} \zeta(\boldsymbol{\theta}^*)}{\|V_0 \boldsymbol{\gamma}\|}\right\} &= \sum\nolimits_{i=1}^n \log I\!\!E \exp\left\{\lambda \sigma_i \frac{\boldsymbol{\gamma}^\top \Psi_i}{\|V_0 \boldsymbol{\gamma}\|} \frac{(Y_i - I\!\!E Y_i)}{\sigma_i}\right\} \\
&\leq \frac{\lambda^2 \nu_0^2}{2} \sum\nolimits_{i=1}^n \sigma_i^2 \frac{\bigl(\boldsymbol{\gamma}^\top \Psi_i\bigr)^2}{\|V_0 \boldsymbol{\gamma}\|^2} \quad \Bigl(\text{for } |\lambda| \leq \sqrt{N_\sigma}\overline{\mathtt{g}}\Bigr) \\
&\leq \lambda^2 \nu_0^2 / 2.
\end{aligned}$$

Table A.3: The GLM case

Source of the values	Corresponding values
$(\boldsymbol{ED_0}), (\boldsymbol{ED_2})$	$\mathtt{g} = \overline{\mathtt{g}}\sqrt{N_\sigma}, \quad \mathtt{g}_2(\mathtt{r}) = +\infty, \quad \omega = 0,$
$(\boldsymbol{\mathcal{L}_0})$	$\delta(\mathtt{r}) = \mathtt{C}_{h_0}\mathtt{r}/\sqrt{N_h},$
$(\boldsymbol{\mathcal{I}})$	$\mathfrak{a}^2 \stackrel{\text{def}}{=} \max\limits_{1\le i\le n}\Big\{\sigma_i^2/h''(\Psi_i^\top\boldsymbol{\theta}^*), \text{for } h''(\Psi_i^\top\boldsymbol{\theta}^*) > 0\Big\},$
$(\boldsymbol{\mathcal{L}\mathtt{r}})$, Th. A.1, A.2, A.5	$\mathtt{b}(\mathtt{r}) = \mathtt{C}_h(\mathtt{r}), \quad \mathfrak{Z}(\mathtt{x}) = \mathtt{C}\sqrt{p+\mathtt{x}}, \quad \mathtt{r}_0 \ge \mathtt{C}\sqrt{p+\mathtt{x}},$
Th. A.3	$\mathfrak{Z}^2_{\mathtt{qf}}(\mathtt{x}, I\!B) \le \mathtt{C}\mathfrak{a}^2(p+6\mathtt{x}),$
Th. A.6	$\mathfrak{Z}^2_{\mathtt{qf}}(\mathtt{x}, \mathcal{B}) \le \mathtt{C}\mathfrak{a}^{\circ 2}(p+6\mathtt{x}) \quad \text{for } \mathfrak{a}^{\circ 2} \text{ from (A.27)},$
Th. A.2	$\Delta_{\mathrm{W}}(\mathtt{r}_0, \mathtt{x}) \le \mathtt{C}\dfrac{p+\mathtt{x}}{\sqrt{N_h}}, \quad \Delta_{\mathrm{W}^2}(\mathtt{r}_0, \mathtt{x}) \le \mathtt{C}\dfrac{(p+\mathtt{x})^{3/2}}{\sqrt{N_h}},$
Th. A.4	$\Delta^\circ_{\mathrm{W}}(\mathtt{r}_0, \mathtt{x}) \le \mathtt{C}\dfrac{p+\mathtt{x}}{\sqrt{N_h}}\Big(1+\sqrt{\dfrac{n}{N_h}}\Big),$
	$\Delta^\circ_{\mathrm{W}^2}(\mathtt{r}_0, \mathtt{x}) \le \mathtt{C}\dfrac{(p+\mathtt{x})^{3/2}}{\sqrt{N_h}}\Big(1+\sqrt{\dfrac{n}{N_h}}\Big),$
Th. A.4, L. A.2	$\omega_1(\mathtt{r}) \le \mathtt{C}\dfrac{n}{N_h}\Big\{\mathtt{C}_{d_0}(\mathtt{r})\dfrac{\mathtt{r}}{\sqrt{N_h n}} + \dfrac{1}{\sqrt{n}}\Big\},$
	$\Delta^\circ_{\xi}(\mathtt{r}_0, \mathtt{x}) \le \mathtt{C}\dfrac{p+\mathtt{x}}{\sqrt{N_h}}\sqrt{\dfrac{n}{N_h}},$
$(\boldsymbol{\mathcal{I}_B})$	$\mathfrak{a}^2_B = \max\limits_{1\le i\le n}\dfrac{\big\{I\!E Y_i - h'(\Psi_i^\top\boldsymbol{\theta}^*)\big\}^2}{h''(\Psi_i^\top\boldsymbol{\theta}^*)},$
$(\boldsymbol{\mathcal{L}_{0m}})$	$\mathtt{C}_m(\mathtt{r}) \le \dfrac{n}{N_h}\Big\{\mathtt{C}_{h_0}(\mathtt{r})\dfrac{\mathtt{r}}{\sqrt{N_h}} + \mathtt{C}\Big\},$
(SmB)	$\delta^2_{\mathrm{smb}} = 1 - \min\limits_{1\le i\le n}\dfrac{\mathrm{Var}\, Y_i}{\mathrm{Var}\, Y_i + \big\{I\!E Y_i - h'(\Psi_i^\top\boldsymbol{\theta}^*)\big\}^2}.$

Due to (A.32) $\nabla^2_{\boldsymbol{\theta}}\zeta(\boldsymbol{\theta}) \equiv \nabla^2_{\boldsymbol{\theta}}\zeta_i(\boldsymbol{\theta}) \equiv 0$, hence conditions $(\boldsymbol{ED_2})$, $(\boldsymbol{ED_{2m}})$ are fulfilled with $\omega = 0$ and an arbitrary large $\mathtt{g}_2(\mathtt{r})$. Now let us check condition $(\boldsymbol{\mathcal{L}_0})$. By (A.34) we have for $\mathtt{r} \le \mathtt{r}_0$, $\forall\, \boldsymbol{\theta} \in \Theta_0(\mathtt{r})$ and some $\overline{\boldsymbol{\theta}} \in \Theta_0(\mathtt{r})$

$$\begin{aligned}
\| D_0^{-1} D^2(\boldsymbol{\theta}) D_0^{-1} - \boldsymbol{I}_p \| &\le \max_{\substack{1 \le i \le n, \\ h''(\Psi_{i,k}^\top \boldsymbol{\theta}^*) > 0}} \left| \frac{h''(\Psi_i^\top \boldsymbol{\theta}) - h''(\Psi_i^\top \boldsymbol{\theta}^*)}{h''(\Psi_i^\top \boldsymbol{\theta}^*)} \right| \\
&\le \mathtt{r} \max_{\substack{1 \le i \le n, \\ h''(\Psi_{i,k}^\top \boldsymbol{\theta}^*) > 0}} \frac{\left| h'''(\Psi_i^\top \overline{\boldsymbol{\theta}}) \right|}{\left| h''(\Psi_i^\top \boldsymbol{\theta}^*) \right|^2} \left\| D_0^{-1} \Psi_i h''(\Psi_i^\top \boldsymbol{\theta}^*) \right\| \\
&\le \mathtt{C}_{h_0} \frac{\mathtt{r}}{\sqrt{N_h}} \quad \text{by condition } \mathbf{(h_0 \,\&\, GLM)}. \qquad \text{(A.35)}
\end{aligned}$$

By definitions (A.33), (A.34) condition $(\boldsymbol{\mathcal{I}})$ is fulfilled with

$$\mathfrak{a}^2 \stackrel{\text{def}}{=} \max_{\substack{1 \le i \le n, \\ h''(\Psi_i^\top \boldsymbol{\theta}^*) > 0}} \left\{ \sigma_i^2 / h''(\Psi_i^\top \boldsymbol{\theta}^*) \right\}. \qquad \text{(A.36)}$$

Below we proceed with condition $(\boldsymbol{\mathcal{L}}\mathtt{r})$. By (A.31) and definition of $\boldsymbol{\theta}^*$:

$$\sum\nolimits_{i=1}^n \Psi_i \left\{ I\!\!E Y_i - h'(\Psi_i^\top \boldsymbol{\theta}^*) \right\} = 0,$$

therefore, by Taylor formula and (A.34) it holds $\forall \boldsymbol{\theta} \in \Theta : \| D_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \| = \mathtt{r}$ and for some $\overline{\boldsymbol{\theta}} \in \Theta_0(\mathtt{r})$:

$$\begin{aligned}
-2\left\{ I\!\!E L(\boldsymbol{\theta}) - I\!\!E L(\boldsymbol{\theta}^*) \right\} &= -2 \sum_{i=1}^n I\!\!E Y_i \Psi_i^\top (\boldsymbol{\theta} - \boldsymbol{\theta}^*) + 2 \sum_{i=1}^n \left\{ h(\Psi_i^\top \boldsymbol{\theta}) - h(\Psi_i^\top \boldsymbol{\theta}^*) \right\} \\
&= -2 (\boldsymbol{\theta} - \boldsymbol{\theta}^*) \sum_{i=1}^n \Psi_i h'(\Psi_i^\top \boldsymbol{\theta}^*) + 2 \sum_{i=1}^n \left\{ h(\Psi_i^\top \boldsymbol{\theta}) - h(\Psi_i^\top \boldsymbol{\theta}^*) \right\} \\
&= (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \sum_{i=1}^n \Psi_i \Psi_i^\top h''(\Psi_i^\top \overline{\boldsymbol{\theta}}) (\boldsymbol{\theta} - \boldsymbol{\theta}^*) \\
&\ge \mathtt{r}^2 \| D_0^{-1} D^2(\overline{\boldsymbol{\theta}}) D_0^{-1} \| \\
&\ge \mathtt{r}^2 \mathtt{C}_h(\mathtt{r}) \text{ by } \mathbf{(h_r \,\&\, GLM)}.
\end{aligned}$$

By definition (A.5) and the obtained above $\mathtt{g} = \overline{\mathtt{g}}\sqrt{N_\sigma}$ it holds $\mathfrak{Z}(\mathtt{x}) = \mathtt{C}\sqrt{p + \mathtt{x}}$. Similarly to the IID case $\mathfrak{Z}^2_{\mathtt{qf}}(\mathtt{x}, I\!\!B) \le \mathtt{C}\mathfrak{a}^2(p + 6\mathtt{x})$, therefore, the concentration condition (A.2) is fulfilled with $\mathtt{r}_0 \le \mathtt{C}\sqrt{p + \mathtt{x}}$. By definitions (A.3), (A.4) it holds

$$\begin{aligned}
\Delta_{\mathrm{W}}(\mathtt{r}_0, \mathtt{x}) &= 3\mathtt{r}_0 \delta(\mathtt{r}_0) \le \mathtt{C} \frac{\mathtt{r}_0^2}{\sqrt{N_h}} \le \mathtt{C} \frac{p + \mathtt{x}}{\sqrt{N_h}}, \\
\Delta_{\mathrm{W}^2}(\mathtt{r}_0, \mathtt{x}) &\le \mathtt{C} \frac{(p + \mathtt{x})^{3/2}}{\sqrt{N_h}}.
\end{aligned} \qquad \text{(A.37)}$$

Now let us consider conditions from Section 3.4.2. Similarly to (A.35)

$$\mathtt{C}_m(\mathtt{r}) \le \frac{n}{N_h}\left\{\mathtt{C}_{d_0}(\mathtt{r})\frac{\mathtt{r}}{\sqrt{N_h}} + \mathtt{C}\right\} \le \mathtt{C}\frac{n}{N_h}\left\{\frac{\mathtt{r}_0}{\sqrt{N_h}} + 1\right\}$$

for $\mathtt{r} \le \mathtt{r}_0$. Hence by (A.18)

$$\omega_1(\mathtt{r}) \le \mathtt{C}\frac{n}{N_h}\left\{\mathtt{C}_{d_0}(\mathtt{r})\frac{\mathtt{r}}{\sqrt{N_h n}} + \frac{1}{\sqrt{n}}\right\},$$

and $\Delta^{\circ}_{\mathrm{W}}(\mathtt{r}_0, \mathtt{x})$, $\Delta^{\circ}_{\mathrm{W}^2}(\mathtt{r}_0, \mathtt{x})$ are bounded similarly to (A.37). By definitions (A.34) and (1.10) conditions $(\boldsymbol{\mathcal{I}_B})$, **(SmB)** are fulfilled with

$$\begin{aligned}
\mathfrak{a}_B^2 &\stackrel{\text{def}}{=} \max_{\substack{1\le i\le n,\\ h''(\Psi_i^\top\boldsymbol{\theta}^*)>0}} \left\{\left[I\!\!E Y_i - h'(\Psi_i^\top\boldsymbol{\theta}^*)\right]^2 / h''(\Psi_i^\top\boldsymbol{\theta}^*)\right\},\\
\delta^2_{\mathrm{smb}} &\stackrel{\text{def}}{=} 1 - \min_{1\le i\le n} \frac{\operatorname{Var} Y_i}{I\!\!E\left\{Y_i - h'(\Psi_i^\top\boldsymbol{\theta}^*)\right\}^2}\\
&= 1 - \min_{1\le i\le n} \frac{\operatorname{Var} Y_i}{\operatorname{Var} Y_i + \left\{I\!\!E Y_i - h'(\Psi_i^\top\boldsymbol{\theta}^*)\right\}^2}.
\end{aligned}$$

Condition $(\boldsymbol{SD_1})$ is implied by the following bound:

$$\begin{aligned}
\frac{1}{N_H}\left\{\max_{1\le i\le n}\frac{\left\{Y_i - h'(\Psi_i^\top\boldsymbol{\theta}^*)\right\}^2}{I\!\!E\left\{Y_i - h'(\Psi_i^\top\boldsymbol{\theta}^*)\right\}^2} - 1\right\} &\le \delta_v^2,\\
\frac{1}{\sqrt{N_H}} &\stackrel{\text{def}}{=} \max_{1\le i\le n}\left\|H_0^{-1}\Psi_i\right\|\sqrt{I\!\!E\left\{Y_i - h'(\Psi_i^\top\boldsymbol{\theta}^*)\right\}^2} \le 1.
\end{aligned}$$

□

A.3.3 Linear quantile regression (QR)

Let the independent observations $Y_1, \ldots, Y_n$ be scalar, and the design points $X_1, \ldots, X_n$ be deterministic. Let $\tau \in (0,1)$ denote a fixed known quantile level. The object of estimation is a quantile function $q_\tau(x)$ s.t.

$$I\!\!P\big(Y_i < q_\tau(X_i)\big) = \tau \quad \forall\, i = 1, \ldots, n.$$

Using the quantile regression approach by Koenker and Bassett Jr (1978), this problem can be treated with quasi maximum likelihood method and the following log-likelihood function:

$$\begin{aligned}
L(\boldsymbol{\theta}) &= -\sum\nolimits_{i=1}^n \rho_\tau\left(Y_i - g(X_i, \boldsymbol{\theta})\right),\\
\rho_\tau(x) &\stackrel{\text{def}}{=} x\left(\tau - \mathbb{1}\{x < 0\}\right),
\end{aligned} \tag{A.38}$$

where $g(\cdot, \boldsymbol{\theta})$ is some known regression function. This log-likelihood function corresponds to asymmetric Laplace distribution with the density $\tau(1-\tau)\mathrm{e}^{-\rho_\tau(x-a)}$. We

consider the linear w.r.t. $\boldsymbol{\theta}$ regression function $g(X_i, \boldsymbol{\theta}) \stackrel{\text{def}}{=} \Psi_i^\top \boldsymbol{\theta}$ for some known deterministic regressors $\Psi_i \in \mathbb{R}^p$. Let $f_i(\boldsymbol{\theta})$ denote the probability density function of Y_i evaluated at the point $\Psi_i^\top \boldsymbol{\theta}$, and $P_i(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \mathbb{P}\{Y_i - \Psi_i^\top \boldsymbol{\theta} < 0\}$.

Lemma A.5. *Consider the linear quantile regression model. If the conditions below are fulfilled:*

($\mathbf{P_{\theta^*}}$ & QR) *For some constant* $\mathtt{C} \geq 1$ *it holds*

$$\mathtt{C}_{P^*} \stackrel{\text{def}}{=} \max_{1 \leq i \leq n} \frac{1}{4P_i(\boldsymbol{\theta}^*)(1 - P_i(\boldsymbol{\theta}^*))} \leq \mathtt{C}, \quad \mathtt{C}_{f^*} \stackrel{\text{def}}{=} \max_{1 \leq i \leq n} \frac{1}{f_i(\boldsymbol{\theta}^*)} \leq \mathtt{C}.$$

($\mathbf{f_0}$ & QR) *For each* $\mathtt{r} \geq 0$ *there exists a value* $\mathtt{C}_{f_0}(\mathtt{r}) \geq 0$ *such that* $\forall\, \boldsymbol{\theta} \in \Theta_0(\mathtt{r})$ *and* $i = 1, \ldots, n$ *it holds:*

$$\left| \frac{f_i'(\boldsymbol{\theta})}{f_i^2(\boldsymbol{\theta}^*)} \right| \leq \mathtt{C}_{f_0}(\mathtt{r}),$$

and $\mathtt{C}_{f_0}(\mathtt{r}) \leq \mathtt{C}_{f_0} = const$ *for* $\mathtt{r} \leq \mathtt{r}_0$.

($\mathbf{f_r}$ & QR) *For each* $\mathtt{r} > \mathtt{r}_0$ *there exists a value* $\mathtt{C}_{f_\mathtt{r}}(\mathtt{r}) > 0$ *s.t.* $\mathtt{r}\mathtt{C}_{f_\mathtt{r}}(\mathtt{r}) \to +\infty$ *with* $\mathtt{r} \to +\infty$, *and* $\forall\, \boldsymbol{\theta} \in \Theta_0(\mathtt{r})$ *it holds*

$$\min_{1 \leq i \leq n} \left| \frac{f_i(\boldsymbol{\theta})}{f_i(\boldsymbol{\theta}^*)} \right| \geq \mathtt{C}_{f_\mathtt{r}}(\mathtt{r}).$$

Then the statements from Sections A.1 and A.2 and their conditions from Section 2.5 depend on the values from Table A.4 with

$$\frac{1}{\sqrt{N_{f^*}}} \stackrel{\text{def}}{=} \max_{1 \leq i \leq n} \| D_0^{-1} \Psi_i f_i(\boldsymbol{\theta}^*) \| \leq 1. \tag{A.39}$$

Proof of Lemma A.5.

$$\nabla_{\boldsymbol{\theta}} \rho_\tau\big(Y_i - \Psi_i^\top \boldsymbol{\theta}\big) = -\Psi_i\big(\tau - \mathbb{1}\{Y_i - \Psi_i^\top \boldsymbol{\theta} < 0\}\big),$$
$$\nabla^2_{\boldsymbol{\theta}} \rho_\tau\big(Y_i - \Psi_i^\top \boldsymbol{\theta}\big) = \Psi_i \Psi_i^\top\, \mathbb{1}\{Y_i - \Psi_i^\top \boldsymbol{\theta} = 0\}.$$

Therefore, by definitions (A.38) it holds

$$\nabla_{\boldsymbol{\theta}} \zeta(\boldsymbol{\theta}) = -\sum\nolimits_{i=1}^n \Psi_i\big(\mathbb{1}\{Y_i - \Psi_i^\top \boldsymbol{\theta} < 0\} - \mathbb{P}\{Y_i - \Psi_i^\top \boldsymbol{\theta} < 0\}\big)$$
$$\nabla^2_{\boldsymbol{\theta}} \zeta(\boldsymbol{\theta}) = -\sum\nolimits_{i=1}^n \Psi_i \Psi_i^\top \big[\mathbb{1}\{Y_i - \Psi_i^\top \boldsymbol{\theta} = 0\} - f_i(\boldsymbol{\theta})\big],$$

and

$$\begin{aligned} V_0^2 &\stackrel{\text{def}}{=} \operatorname{Var} \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}^*) = \sum\nolimits_{i=1}^n \Psi_i \Psi_i^\top P_i(\boldsymbol{\theta}^*)(1 - P_i(\boldsymbol{\theta}^*)), \\ D^2(\boldsymbol{\theta}) &\stackrel{\text{def}}{=} -\nabla^2_{\boldsymbol{\theta}} \mathbb{E} L(\boldsymbol{\theta}) = \sum\nolimits_{i=1}^n \Psi_i \Psi_i^\top f_i(\boldsymbol{\theta}). \end{aligned} \tag{A.40}$$

Table A.4: Linear quantile regression

Source of the values	Corresponding values
$(\boldsymbol{ED_0})$, $(\boldsymbol{ED_2})$	$\mathtt{g} = \mathtt{g}_2(\mathtt{r}) = +\infty, \quad \omega = \mathtt{C}_{f^*}/(2\sqrt{N_{f^*}})$,
$(\boldsymbol{\mathcal{L}_0})$	$\delta(\mathtt{r}) = \mathtt{C}_{f_0}\mathtt{r}/\sqrt{N_{f^*}}$,
$(\boldsymbol{\mathcal{I}})$	$\mathfrak{a}^2 = \max\limits_{1\le i\le n} \dfrac{P_i(\boldsymbol{\theta}^*)(1-P_i(\boldsymbol{\theta}^*))}{f_i(\boldsymbol{\theta}^*)} \le \dfrac{\mathtt{C}_{f^*}}{4}$,
$(\boldsymbol{\mathcal{L}\mathtt{r}})$, Th. A.1, A.5	$\mathtt{b}(\mathtt{r}) = \mathtt{C}_{f_r}(\mathtt{r}), \quad \mathfrak{Z}(\mathtt{x}) = \mathtt{C}\sqrt{p+\mathtt{x}}, \quad \mathtt{r}_0 \ge \mathtt{C}\sqrt{p+\mathtt{x}}$,
Th. A.3	$\mathfrak{Z}^2_{\mathtt{qf}}(\mathtt{x}, I\!B) \le \mathtt{C}\mathfrak{a}^2(p+6\mathtt{x})$,
Th. A.6	$\mathfrak{Z}^2_{\mathtt{qf}}(\mathtt{x}, \mathcal{B}) \le \mathtt{C}\mathfrak{a}^{\circ 2}(p+6\mathtt{x})$ for $\mathfrak{a}^{\circ 2}$ from (A.27),
Th. A.2	$\Delta_{\mathrm{W}}(\mathtt{r}_0, \mathtt{x}) \le \mathtt{C}\dfrac{p+\mathtt{x}}{\sqrt{N_{f^*}}}, \quad \Delta_{\mathrm{W}^2}(\mathtt{r}_0, \mathtt{x}) \le \mathtt{C}\dfrac{(p+\mathtt{x})^{3/2}}{\sqrt{N_{f^*}}}$,
Th. A.4	$\Delta^{\circ}_{\mathrm{W}}(\mathtt{r}_0, \mathtt{x}) \le \mathtt{C}\dfrac{p+\mathtt{x}}{\sqrt{N_{f^*}}}\Big(\sqrt{\mathtt{x}} + \sqrt{\dfrac{n}{N_{f^*}}}\Big)$,
	$\Delta^{\circ}_{\mathrm{W}^2}(\mathtt{r}_0, \mathtt{x}) \le \mathtt{C}\dfrac{(p+\mathtt{x})^{3/2}}{\sqrt{N_{f^*}}}\Big(\sqrt{\mathtt{x}} + \sqrt{\dfrac{n}{N_{f^*}}}\Big)$,
Th. A.4	$\omega_1(\mathtt{r}, \mathtt{x}) \le \mathtt{C}\dfrac{n}{N_{f^*}}\Big\{\mathtt{C}_{f_0}(\mathtt{r})\dfrac{\mathtt{r}}{\sqrt{nN_{f^*}}} + \dfrac{1}{\sqrt{n}}\Big\} + \dfrac{\sqrt{\mathtt{x}}\mathtt{C}_{f^*}}{\sqrt{N_{f^*}}}$,
L. A.2	$\Delta^{\circ}_{\xi}(\mathtt{r}_0, \mathtt{x}) \le \mathtt{C}\dfrac{p+\mathtt{x}}{\sqrt{N_{f^*}}}\Big(\sqrt{\mathtt{x}} + \sqrt{n/N_{f^*}}\Big)$,
$(\boldsymbol{\mathcal{L}_{0m}})$	$\mathtt{C}_m(\mathtt{r}) \le \dfrac{n}{N_{f^*}}\Big\{\mathtt{C}_{f_0}(\mathtt{r})\dfrac{\mathtt{r}}{\sqrt{N_{f^*}}} + \mathtt{C}\Big\}$,
$(\boldsymbol{\mathcal{I}_B})$	$\mathfrak{a}^2_B = \max\limits_{1\le i\le n} \dfrac{\big(\tau - I\!P\big\{Y_i - \Psi_i^\top\boldsymbol{\theta}^* < 0\big\}\big)^2}{f_i(\boldsymbol{\theta}^*)}$,
(SmB)	$\delta^2_{\mathrm{smb}} = 1 - \min\limits_{1\le i\le n} \dfrac{\mathrm{Var}\left(\tau - \mathbb{1}\big\{Y_i - \Psi_i^\top\boldsymbol{\theta}^* < 0\big\}\right)}{\mathrm{Var}\left(\tau - \mathbb{1}\big\{Y_i - \Psi_i^\top\boldsymbol{\theta}^* < 0\big\}\right) + \big(\tau - I\!P\big\{Y_i - \Psi_i^\top\boldsymbol{\theta}^* < 0\big\}\big)^2}$.

Condition $(\boldsymbol{ED_0})$ reads as follows:

$$\log I\!\!E \exp\left\{\lambda \frac{\boldsymbol{\gamma}^\top \nabla_{\boldsymbol{\theta}} \zeta(\boldsymbol{\theta}^*)}{\|V_0 \boldsymbol{\gamma}\|}\right\}$$
$$= \sum_{i=1}^n -P_i(\boldsymbol{\theta}^*) \frac{\lambda \boldsymbol{\gamma}^\top \Psi_i}{\|V_0 \boldsymbol{\gamma}\|} + \sum_{i=1}^n \log\left\{1 - P_i(\boldsymbol{\theta}^*) + P_i(\boldsymbol{\theta}^*) \exp\left(\frac{\lambda \boldsymbol{\gamma}^\top \Psi_i}{\|V_0 \boldsymbol{\gamma}\|}\right)\right\}$$
$$\leq \frac{1}{8} \sum_{i=1}^n \left(\frac{\lambda \boldsymbol{\gamma}^\top \Psi_i}{\|V_0 \boldsymbol{\gamma}\|}\right)^2 \leq \frac{1}{8} \lambda^2 \max_{1 \leq i \leq n} \frac{1}{P_i(\boldsymbol{\theta}^*)(1 - P_i(\boldsymbol{\theta}^*))} \leq \mathtt{C}_{P^*} \lambda^2 / 2 \quad \forall\, \lambda \in I\!R.$$

Similarly for the conditions $(\boldsymbol{ED_2})$, $(\boldsymbol{ED_{2m}})$ it holds

$$\log I\!\!E \exp\left\{\frac{\lambda}{\omega} \boldsymbol{\gamma}_1^\top D_0^{-1} \nabla^2_{\boldsymbol{\theta}} \zeta(\boldsymbol{\theta}) D_0^{-1} \boldsymbol{\gamma}_2\right\} \leq \frac{\lambda^2}{8} \frac{\mathtt{C}^2_{f^*}}{N_{f^*} \omega^2} \leq \lambda^2/2$$

for $\omega \geq \mathtt{C}_{f^*}/(2\sqrt{N_{f^*}})$. Condition $(\boldsymbol{\mathcal{L}_0})$ is implied by (A.40), (A.39) and $(\mathbf{f_0}\,\&\,\mathbf{QR})$:

$$\|D_0^{-1} D^2(\boldsymbol{\theta}) D_0^{-1} - \boldsymbol{I}_p\| \leq \max_{1 \leq i \leq n} \left|\frac{f_i(\boldsymbol{\theta}) - f_i(\boldsymbol{\theta}^*)}{f_i(\boldsymbol{\theta}^*)}\right|$$
$$\leq \mathtt{C}_{f_0} \max_{1 \leq i \leq n} \left\|f_i(\boldsymbol{\theta}^*) D_0^{-1} \Psi_i\right\| \|D_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \leq \mathtt{C}_{f_0} \mathtt{r} / \sqrt{N_{f^*}}. \tag{A.41}$$

Due to (A.40) and $(\mathbf{P}_{\boldsymbol{\theta}^*}\,\&\,\mathbf{QR})$ condtion $(\boldsymbol{\mathcal{I}})$ is justified with

$$\mathfrak{a}^2 = \max_{1 \leq i \leq n} \frac{P_i(\boldsymbol{\theta}^*)(1 - P_i(\boldsymbol{\theta}^*))}{f_i(\boldsymbol{\theta}^*)} \leq \frac{\mathtt{C}_{f^*}}{4}.$$

Condition $(\boldsymbol{\mathcal{L}}\mathtt{r})$ is derived straightforwardly from $(\mathbf{f}_\mathtt{r}\,\&\,\mathbf{QR})$:

$$-2\{I\!\!E L(\boldsymbol{\theta}) - I\!\!E L(\boldsymbol{\theta}^*)\} \geq \mathtt{r}^2 \|D_0^{-1} D^2(\overline{\boldsymbol{\theta}}) D_0^{-1}\| \geq \mathtt{r}^2 \mathtt{C}_{f_\mathtt{r}}(\mathtt{r}).$$

By definition (A.5) and the obtained above $\mathtt{g} = +\infty$ it holds $\mathfrak{Z}(\mathtt{x}) = \mathtt{C}\sqrt{p + \mathtt{x}}$. Similarly to the IID case $\mathfrak{Z}^2_{\mathtt{qf}}(\mathtt{x}, I\!B) \leq \mathtt{C}\mathfrak{a}^2(p + 6\mathtt{x})$, therefore, the concentration condition (A.2) is fulfilled with $\mathtt{r}_0 \leq \mathtt{C}\sqrt{p + \mathtt{x}}$. By definitions (A.3), (A.4) it holds

$$\Delta_{\mathrm{W}}(\mathtt{r}_0, \mathtt{x}) \leq \mathtt{C} \frac{p + \mathtt{x}}{\sqrt{N_{f^*}}}, \quad \Delta_{\mathrm{W}^2}(\mathtt{r}_0, \mathtt{x}) \leq \mathtt{C} \frac{(p + \mathtt{x})^{3/2}}{\sqrt{N_{f^*}}}. \tag{A.42}$$

For $(\boldsymbol{\mathcal{L}_{0m}})$ it holds similarly to (A.41)

$$\|D_0^{-1} \nabla^2_{\boldsymbol{\theta}} I\!\!E \ell_i(\boldsymbol{\theta}) D_0^{-1}\| \leq \max_{1 \leq i \leq n} \frac{f_i(\boldsymbol{\theta})}{f_i(\boldsymbol{\theta}^*)} \frac{1}{N_{f^*}} \leq \frac{1}{N_{f^*}} \Bigl(1 + \frac{\mathtt{r}}{\sqrt{N_{f^*}}} \mathtt{C}_{f_0}(\mathtt{r})\Bigr).$$

Hence by (A.18) $\omega_1(\mathtt{r}) \leq \mathtt{C}\frac{n}{N_{f^*}} \left\{\mathtt{C}_{f_0}(\mathtt{r}) \frac{\mathtt{r}}{\sqrt{n N_{f^*}}} + \frac{1}{\sqrt{n}}\right\} + \frac{\sqrt{\mathtt{x}}\mathtt{C}_{f^*}}{\sqrt{N_{f^*}}}$. Conditions $(\boldsymbol{\mathcal{I}_B})$, **(SmB)** are fulfilled with

$$\mathfrak{a}^2_B \stackrel{\mathrm{def}}{=} \max_{1 \leq i \leq n} \frac{\bigl(\tau - I\!P\bigl\{Y_i - \Psi_i^\top \boldsymbol{\theta}^* < 0\bigr\}\bigr)^2}{f_i(\boldsymbol{\theta}^*)},$$
$$\delta^2_{\mathrm{smb}} \stackrel{\mathrm{def}}{=} 1 - \min_{1 \leq i \leq n} \frac{\mathrm{Var}\bigl(\tau - I\!\!1\bigl\{Y_i - \Psi_i^\top \boldsymbol{\theta}^* < 0\bigr\}\bigr)}{\mathrm{Var}\bigl(\tau - I\!\!1\bigl\{Y_i - \Psi_i^\top \boldsymbol{\theta}^* < 0\bigr\}\bigr) + \bigl(\tau - I\!P\bigl\{Y_i - \Psi_i^\top \boldsymbol{\theta}^* < 0\bigr\}\bigr)^2}.$$

Condition $(\boldsymbol{SD_1})$ is implied by the following bound:

$$I\!P\left(\max_{1\le i\le n}\left\{\frac{\mathbb{1}\{Y_i-\Psi_i^\top\boldsymbol{\theta}^*<0\}-I\!P\{Y_i-\Psi_i^\top\boldsymbol{\theta}^*<0\}}{I\!E\left\{Y_i-h'(\Psi_i^\top\boldsymbol{\theta}^*)\right\}^2}\right\}c_i(\tau)\le\delta_v^2N_{H_\tau}\right)\ge 1-\mathrm{e}^{-\mathtt{x}},$$

$$c_i(\tau)\stackrel{\text{def}}{=}\frac{(1-2\tau)}{\tau^2-(1-2\tau)I\!P\{Y_i-\Psi_i^\top\boldsymbol{\theta}^*<0\}},$$

$$\frac{1}{\sqrt{N_{H_\tau}}}\stackrel{\text{def}}{=}\max_{1\le i\le n}\|H_0^{-1}\Psi_i\|\sqrt{I\!E\left(\tau-\mathbb{1}\{Y_i-\Psi_i^\top\boldsymbol{\theta}^*<0\}\right)^2}\le 1.$$

□

A.3.4 Small modelling bias condition for some models

Table A.5 collects the bounds on the value $\|H_0^{-1}B_0^2H_0^{-1}\|$ from condition **(SmB)**, obtained above for some models. In the IID case $\nabla_{\boldsymbol{\theta}}I\!E\ell_i(\boldsymbol{\theta}^*)\equiv 0$, therefore, $B_0^2=0$. For the GLM

$$\|H_0^{-1}B_0^2H_0^{-1}\|\le 1-\min_{1\le i\le n}\frac{\operatorname{Var}Y_i}{\operatorname{Var}Y_i+\left\{I\!EY_i-h'(\Psi_i^\top\boldsymbol{\theta}^*)\right\}^2}\in[0,1).$$

It is important that $I\!E_{\boldsymbol{\theta}^*}Y_i=h'(\Psi_i^\top\boldsymbol{\theta}^*)$, i.e. in the case of the true parametric model $I\!P\in\{I\!P_\upsilon\}$ the modelling bias is indeed equal to zero. For the quantile regression model the bound is similar:

$$\|H_0^{-1}B_0^2H_0^{-1}\|\le 1-\min_{1\le i\le n}\frac{\operatorname{Var}\left(\tau-\mathbb{1}\{Y_i-\Psi_i^\top\boldsymbol{\theta}^*<0\}\right)}{\operatorname{Var}\left(\tau-\mathbb{1}\{Y_i-\Psi_i^\top\boldsymbol{\theta}^*<0\}\right)+\left(\tau-I\!P\{Y_i-\Psi_i^\top\boldsymbol{\theta}^*<0\}\right)^2}.$$

If $I\!P\{Y_i-\Psi_i^\top\boldsymbol{\theta}^*<0\}\equiv\tau$, then the right side of the last inequality is equal to zero.

Table A.5: The modelling bias for some models

Model	δ^2_{smb}
IID, A.3.1	0
GLM, A.3.2	$1-\min\limits_{1\le i\le n}\dfrac{\operatorname{Var}Y_i}{\operatorname{Var}Y_i+\left\{I\!EY_i-h'(\Psi_i^\top\boldsymbol{\theta}^*)\right\}^2}$
QR, A.3.3	$1-\min\limits_{1\le i\le n}\dfrac{\operatorname{Var}\left(\tau-\mathbb{1}\{Y_i-\Psi_i^\top\boldsymbol{\theta}^*<0\}\right)}{\operatorname{Var}\left(\tau-\mathbb{1}\{Y_i-\Psi_i^\top\boldsymbol{\theta}^*<0\}\right)+\left(\tau-I\!P\{Y_i-\Psi_i^\top\boldsymbol{\theta}^*<0\}\right)^2}$

A.4 Simultaneous square-root Wilks approximations

Here we restate the results of Sections A.1 and A.2 for the problem of constructing the simultaneous confidence sets. Let us previously introduce some necessary objects. For each $k = 1, \ldots, K$ $\Theta_{0,k}(\mathtt{r})$ denotes the elliptic vicinity around the true point $\boldsymbol{\theta}_k^*$:

$$\Theta_{0,k}(\mathtt{r}) \stackrel{\text{def}}{=} \{\boldsymbol{\theta} \in \Theta_k : \|D_k(\boldsymbol{\theta} - \boldsymbol{\theta}_k^*)\| \leq \mathtt{r}\}, \tag{A.43}$$

D_k^2 denotes the full Fisher information $p_k \times p_k$ matrix, which is deterministic, symmetric and positive-definite:

$$D_k^2 \stackrel{\text{def}}{=} -\nabla_{\boldsymbol{\theta}}^2 I\!\!E L_k(\boldsymbol{\theta}_k^*).$$

Matrices $I\!\!B_k \stackrel{\text{def}}{=} D_k^{-1} V_k^2 D_k^{-1}$, $\mathcal{B}_k^2 \stackrel{\text{def}}{=} D_k^{-1} \mathcal{V}^2(\boldsymbol{\theta}_k^*) D_k^{-1}$ for $\mathcal{V}_k^2(\boldsymbol{\theta}_k^*) \stackrel{\text{def}}{=} \mathrm{Var}^{\circ}\, \nabla_{\boldsymbol{\theta}} L_k^{\circ}(\boldsymbol{\theta}_k^*)$ are analogous to $I\!\!B = D_0^{-1} V_0^2 D_0^{-1}$ and $\mathcal{B}^2 \stackrel{\text{def}}{=} D_0^{-1} \mathcal{V}^2(\boldsymbol{\theta}^*) D_0^{-1}$, $\mathcal{V}^2(\boldsymbol{\theta}^*) \stackrel{\text{def}}{=} \mathrm{Var}^{\circ}\, \nabla_{\boldsymbol{\theta}} L^{\circ}(\boldsymbol{\theta}^*$ from Theorems (A.3) and (A.6). $\boldsymbol{\xi}_k$ denotes the normalised score:

$$\boldsymbol{\xi}_k \stackrel{\text{def}}{=} D_k^{-1} \nabla_{\boldsymbol{\theta}} L_k(\boldsymbol{\theta}_k^*).$$

Introduce also the following objects similarly to (A.3)-(A.5) and (A.18)

$$\begin{aligned}
\Delta_{k,\mathrm{W}}(\mathtt{r}, \mathtt{x}) &\stackrel{\text{def}}{=} 3\mathtt{r} \left\{ \delta_k(\mathtt{r}) + 6\nu_k \, \mathfrak{Z}_k(\mathtt{x}) \omega_k \right\}, \\
\Delta_{k,\mathrm{W}^2}(\mathtt{r}, \mathtt{x}) &\stackrel{\text{def}}{=} \frac{2}{3} \left\{ 2\mathtt{r} + \mathfrak{Z}_{\mathtt{qf}}(\mathtt{x}, I\!\!B_k) \right\} \Delta_{k,\mathrm{W}}(\mathtt{r}, \mathtt{x}), \\
\mathfrak{Z}_k(\mathtt{x}) &\stackrel{\text{def}}{=} 2\sqrt{p_k} + \sqrt{2\mathtt{x}} + 4p_k(\mathtt{x}\mathtt{g}_k^{-2} + 1)\mathtt{g}_k^{-1}, \\
\omega_{1,k} &\stackrel{\text{def}}{=} \frac{\mathtt{C}_{m,k}(\mathtt{r})}{\sqrt{n}} + 2\omega_k \nu_k \sqrt{2\mathtt{x}}.
\end{aligned}$$

Lemma A.6 (Simultaneous concentration bounds)**.**

1. *If the conditions* $(\boldsymbol{ED_{0,k}})$, $(\boldsymbol{ED_{2,k}})$, $(\boldsymbol{\mathcal{L}_{0,k}})$, $(\boldsymbol{\mathcal{I}_k})$ *and* $(\boldsymbol{\mathcal{L}\mathtt{r}_k})$ *are fulfilled, and for each* $k = 1, \ldots, K$ *the inequality* (A.44) *holds for the constants* $\mathtt{r}_{0,k} > 0$ *and for the functions* $\mathtt{b}_k(\mathtt{r})$ *from* $(\boldsymbol{\mathcal{L}\mathtt{r}_k})$*:*

$$\mathtt{b}_k(\mathtt{r})\mathtt{r} \geq 2 \left\{ \mathfrak{Z}_{\mathtt{qf}}(\mathtt{x}, I\!\!B_k) + 6\omega_k \nu_k \, \mathfrak{Z}_k(\mathtt{x} + \log(2\mathtt{r}/\mathtt{r}_{0,k})) \right\}, \quad \mathtt{r} > \mathtt{r}_{0,k}, \tag{A.44}$$

where $\mathtt{x} = \mathtt{x}_1 + \log(K)$ *for some* $\mathtt{x}_1 > 0$*, then*

$$I\!\!P\left(\bigcup\nolimits_{k=1}^{K} \left\{ \widetilde{\boldsymbol{\theta}}_k \notin \Theta_{0,k}(\mathtt{r}_{0,k}) \right\} \right) \leq 3\mathrm{e}^{-\mathtt{x}_1}.$$

The constants ω_k, ν_k *and* $\mathfrak{a}_k$ *come from the imposed conditions* $(\boldsymbol{ED_{0,k}})$ – $(\boldsymbol{\mathcal{I}_k})$ *(from Section 3.4). In the case A.3.1* $\mathtt{r}_{0,k} \geq \mathtt{C}\sqrt{p_k + \mathtt{x}}$*.*

2. *Let the conditions of the previous part of the lemma be fulfilled. Suppose that the conditions* $(\boldsymbol{Eb})$, $(\widehat{\boldsymbol{SD}_1})$, $(\boldsymbol{\mathcal{I}_{B,k}})$ *hold,* $\mathtt{g}_k \geq \sqrt{2\,\mathrm{tr}(\mathcal{B}_k^2)}$, *and inequality* (A.45) *below holds for each* $k = 1, \ldots, K$ *with* $\mathtt{x} = \mathtt{x}_1 + \log(K)$ *for some* $\mathtt{x}_1 > 0$,

$$\begin{aligned} \mathtt{b}_k(\mathtt{r})\mathtt{r} \geq 2 \bigl\{ \mathfrak{z}_{\mathtt{qf}}(\mathtt{x}, I\!\!B_k) + \mathfrak{z}_{\mathtt{qf}}(\mathtt{x}, \mathcal{B}_k) + 6\nu_k\, \mathfrak{z}_k(\mathtt{x})\omega_{1,k}(\mathtt{r}_{0,k})\mathtt{r}_{0,k} \bigr\} \\ + \; 12\nu_k(\omega_k + \omega_{1,k}(\mathtt{r},\mathtt{x}))\, \mathfrak{z}_k(\mathtt{x} + \log(2\mathtt{r}/\mathtt{r}_{0,k})) \quad \textit{for} \;\; \mathtt{r} > \mathtt{r}_{0,k}, \end{aligned} \tag{A.45}$$

then

$$I\!\!P^\circ \Bigl(\bigcup\nolimits_{k=1}^K \bigl\{ \widetilde{\boldsymbol{\theta}}_k^\circ \notin \Theta_{0,k}(\mathtt{r}_{0,k}) \bigr\} \Bigr) \leq 3\mathrm{e}^{-\mathtt{x}_1}.$$

with $I\!\!P$*-probability* $\geq 1 - 3\mathrm{e}^{-\mathtt{x}_1}$

Lemma A.7 (Simultaneous Wilks approximations)**.**

1. *Let the conditions of part 1 of Lemma A.6 be fulfilled for some* $\mathtt{r}_{0,k} > 0$ *and* $\mathtt{x} = \mathtt{x}_1 + \log(K)$, *then it holds*

$$I\!\!P \Bigl(\bigcap\nolimits_{k=1}^K \bigl\{ \bigl| 2\bigl\{ L_k(\widetilde{\boldsymbol{\theta}}_k) - L_k(\boldsymbol{\theta}_k^*) \bigr\} - \|\boldsymbol{\xi}_k\|^2 \bigr| \leq \Delta_{k,\mathrm{W}^2}(\mathtt{r}_{0,k}, \mathtt{x}_1 + \log(K)) \bigr\} \Bigr) \geq 1 - 5\mathrm{e}^{-\mathtt{x}_1},$$

$$I\!\!P \Bigl(\bigcap\nolimits_{k=1}^K \Bigl\{ \Bigl| \sqrt{2\bigl\{ L_k(\widetilde{\boldsymbol{\theta}}_k) - L_k(\boldsymbol{\theta}_k^*) \bigr\}} - \|\boldsymbol{\xi}_k\| \Bigr| \leq \Delta_{k,\mathrm{W}}(\mathtt{r}_{0,k}, \mathtt{x}_1 + \log(K)) \Bigr\} \Bigr) \geq 1 - 5\mathrm{e}^{-\mathtt{x}_1}.$$

In the case A.3.1 it holds for $\mathtt{r} \leq \mathtt{r}_{0,k}$ *:*

$$\Delta_{k,\mathrm{W}}(\mathtt{r},\mathtt{x}) \leq \mathtt{C}\,\frac{p_k + \mathtt{x}}{\sqrt{n}}, \quad \Delta_{k,\mathrm{W}^2}(\mathtt{r},\mathtt{x}) \leq \mathtt{C}\sqrt{\frac{(p_k+\mathtt{x})^3}{n}}.$$

2. *Let the conditions of parts 1,2 of Lemma A.6 be fulfilled for some* $\mathtt{r}_{0,k} > 0$ *and* $\mathtt{x} = \mathtt{x}_1 + \log(K)$, *then it holds with* $I\!\!P$*-probability* $\geq 1 - 5\mathrm{e}^{-\mathtt{x}_1}$

$$I\!\!P^\circ \Bigl(\bigcap_{k=1}^K \Bigl\{ \Bigl| \sup_{\boldsymbol{\theta}\in\Theta_k} 2\bigl\{ L_k^\circ(\boldsymbol{\theta}) - L_k^\circ(\widetilde{\boldsymbol{\theta}}_k) \bigr\} - \|\boldsymbol{\xi}_k^\circ(\widetilde{\boldsymbol{\theta}}_k)\|^2 \Bigr| \leq \Delta_{k,\mathrm{W}^2}^\circ(\mathtt{r}_{0,k}, \mathtt{x}_1 + \log(K)) \Bigr\} \Bigr) \geq 1 - 4\mathrm{e}^{-\mathtt{x}_1},$$

$$I\!\!P^\circ \Bigl(\bigcap_{k=1}^K \Bigl\{ \Bigl| \sqrt{\sup_{\boldsymbol{\theta}\in\Theta_k} 2\bigl\{ L_k^\circ(\boldsymbol{\theta}) - L_k^\circ(\widetilde{\boldsymbol{\theta}}_k) \bigr\}} - \|\boldsymbol{\xi}_k^\circ(\widetilde{\boldsymbol{\theta}}_k)\| \Bigr| \leq \Delta_{k,\mathrm{W}}^\circ(\mathtt{r}_{0,k}, \mathtt{x}_1 + \log(K)) \Bigr\} \Bigr) \geq 1 - 4\mathrm{e}^{-\mathtt{x}_1}.$$

For the case A.3.1 and $\mathtt{r} \leq \mathtt{r}_{0,k}$ *it holds:*

$$\Delta_{k,\mathrm{W}}^\circ(\mathtt{r},\mathtt{x}) \leq \mathtt{C}\,\frac{p_k + \mathtt{x}}{\sqrt{n}}\sqrt{\mathtt{x}}, \quad \Delta_{k,\mathrm{W}^2}^\circ(\mathtt{r},\mathtt{x}) \leq \mathtt{C}\sqrt{\frac{(p_k+\mathtt{x})^3}{n}}\sqrt{\mathtt{x}}.$$

Lemma A.8. *Let the conditions* $(\boldsymbol{Eb})$, $(\boldsymbol{\mathcal{L}_{0m,k}})$ *and* $(\boldsymbol{ED_{2m,k}})$ *be fulfilled, then for each* $k = 1, \ldots, K$ *it holds for* $\mathtt{r} \le \mathtt{r}_{0,k}$ *with* $I\!P$*-probability* $\ge 1 - \mathrm{e}^{-\mathtt{x}}$

$$I\!P^{\circ}\left(\bigcap\nolimits_{k=1}^{K} \left\{ \sup_{\substack{\boldsymbol{\theta} \in \Theta_{0,k}(\mathtt{r}), \\ \mathtt{r} \le \mathtt{r}_{0,k}}} \| \boldsymbol{\xi}_k^{\circ}(\boldsymbol{\theta}) - \boldsymbol{\xi}_k^{\circ}(\boldsymbol{\theta}_k^*) \| \le \Delta_{\xi,k}^{\circ}(\mathtt{r}, \mathtt{x} + \log(K)) \right\} \right) \ge 1 - \mathrm{e}^{-\mathtt{x}},$$

where

$$\Delta_{\xi,k}^{\circ}(\mathtt{r}, \mathtt{x}) \stackrel{\mathrm{def}}{=} 6\nu_k \, \mathfrak{Z}_k(\mathtt{x}) \omega_{1,k}(\mathtt{r}, \mathtt{x}) \mathtt{r}.$$

In the case A.3.1 it holds for the bounding term

$$\Delta_{\xi,k}^{\circ}(\mathtt{r}_{0,k}, \mathtt{x}) \le \mathtt{C} \frac{p_k + \mathtt{x}}{\sqrt{n}} \sqrt{\mathtt{x}}.$$

Appendix B

Approximation of distributions of ℓ_2-norms

Here we compare probability distributions of ℓ_2-norms of sums of two sets of independent centered vectors: $\boldsymbol{\phi} \stackrel{\text{def}}{=} \sum_{i=1}^n \boldsymbol{\phi}_i$, $\boldsymbol{\psi} \stackrel{\text{def}}{=} \sum_{i=1}^n \boldsymbol{\psi}_i$. Theorem B.1 gives the conditions on the covariance matrices of $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$, and on the 3-d moments $I\!E\|\boldsymbol{\phi}_i\|^3$, $I\!E\|\boldsymbol{\psi}_i\|^3$, $i = 1\ldots, n$ which ensure that the distributions of $\|\boldsymbol{\phi}\|$ and $\|\boldsymbol{\psi}\|$ are close to each other.

Consider two samples $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_n$ and $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_n$, each consists of centered independent random vectors in $I\!R^p$ with nearly the same second moments. This chapter explains how one can quantify the closeness in distribution between the norms of $\boldsymbol{\phi} = \sum_i \boldsymbol{\phi}_i$ and of $\boldsymbol{\psi} = \sum_i \boldsymbol{\psi}_i$. Suppose that

$$I\!E\boldsymbol{\phi}_i = I\!E\boldsymbol{\psi}_i = 0, \quad \operatorname{Var}\boldsymbol{\phi}_i = \Sigma_i, \quad \operatorname{Var}\boldsymbol{\psi}_i = \breve{\Sigma}_i, \quad i = 1, \ldots, n.$$

Let also

$$\boldsymbol{\phi} \stackrel{\text{def}}{=} \sum\nolimits_{i=1}^n \boldsymbol{\phi}_i, \quad \boldsymbol{\psi} \stackrel{\text{def}}{=} \sum\nolimits_{i=1}^n \boldsymbol{\psi}_i, \tag{B.1}$$

$$\Sigma \stackrel{\text{def}}{=} \operatorname{Var}\boldsymbol{\phi} = \sum\nolimits_{i=1}^n \Sigma_i, \quad \breve{\Sigma} \stackrel{\text{def}}{=} \operatorname{Var}\boldsymbol{\psi} = \sum\nolimits_{i=1}^n \breve{\Sigma}_i. \tag{B.2}$$

Introduce also multivariate Gaussian vectors $\overline{\boldsymbol{\phi}}_i, \overline{\boldsymbol{\psi}}_i$ which are mutually independent for $i = 1, \ldots, n$ and

$$\overline{\boldsymbol{\phi}}_i \sim \mathcal{N}(0, \Sigma_i), \quad \overline{\boldsymbol{\psi}}_i \sim \mathcal{N}(0, \breve{\Sigma}_i),$$

$$\overline{\boldsymbol{\phi}} \stackrel{\text{def}}{=} \sum\nolimits_{i=1}^n \overline{\boldsymbol{\phi}}_i \sim \mathcal{N}(0, \Sigma), \quad \overline{\boldsymbol{\psi}} \stackrel{\text{def}}{=} \sum\nolimits_{i=1}^n \overline{\boldsymbol{\psi}}_i \sim \mathcal{N}(0, \breve{\Sigma}). \tag{B.3}$$

The bar sign for a vector stands here for a normal distribution. The following theorem gives the conditions on Σ and $\breve{\Sigma}$ which ensure that $\|\boldsymbol{\phi}\|$ and $\|\boldsymbol{\psi}\|$ are close to each

other in distribution. It also presents a general result on Gaussian approximation of $\|\boldsymbol{\phi}\|$ with $\|\overline{\boldsymbol{\phi}}\|$.

Introduce the following deterministic values, which are supposed to be finite:

$$\delta_n \stackrel{\text{def}}{=} \frac{1}{2}\sum_{i=1}^{n} I\!\!E\left(\|\boldsymbol{\phi}_i\|^3 + \|\overline{\boldsymbol{\phi}}_i\|^3\right), \quad \breve{\delta}_n \stackrel{\text{def}}{=} \frac{1}{2}\sum_{i=1}^{n} I\!\!E\left(\|\boldsymbol{\psi}_i\|^3 + \|\overline{\boldsymbol{\psi}}_i\|^3\right). \tag{B.4}$$

Theorem B.1. *Consider the random vectors $\boldsymbol{\phi}, \boldsymbol{\psi}$ given in* (B.1)*, and their Gaussian analogs defined in* (B.3)*. Assume for the covariance matrices $\Sigma \stackrel{\text{def}}{=} \operatorname{Var}\boldsymbol{\phi}$, $\breve{\Sigma} \stackrel{\text{def}}{=} \operatorname{Var}\boldsymbol{\psi}$ that*

$$\left\|\breve{\Sigma}^{-1/2}\Sigma\breve{\Sigma}^{-1/2} - \boldsymbol{I}_p\right\| \le 1/2, \quad \textit{and} \quad \operatorname{tr}\left\{\left(\breve{\Sigma}^{-1/2}\Sigma\breve{\Sigma}^{-1/2} - \boldsymbol{I}_p\right)^2\right\} \le \delta_\Sigma^2 \tag{B.5}$$

for some $\delta_\Sigma^2 \ge 0$. The sign $\|\cdot\|$ for matrices denotes the spectral norm. Let also $z, \overline{z} \ge \max\{2, \sqrt{p}\}$ and $|z - \overline{z}| \le \delta_z$ for some $\delta_z \ge 0$. Then it holds for all $0 < \Delta \le 0.22$

1.1. $\left|I\!\!P\left(\|\boldsymbol{\phi}\| > z\right) - I\!\!P\left(\|\overline{\boldsymbol{\psi}}\| > \overline{z}\right)\right| \le 16\delta_n\Delta^{-3} + (\Delta + \delta_z)/\sqrt{2} + \delta_\Sigma/2,$

1.2. $\left|I\!\!P\left(\|\boldsymbol{\phi}\| > z\right) - I\!\!P\left(\|\boldsymbol{\psi}\| > \overline{z}\right)\right| \le 16\Delta^{-3}\left(\delta_n + \breve{\delta}_n\right) + (2\Delta + \delta_z)/\sqrt{2} + \delta_\Sigma/2.$

Moreover, if $\max\{\delta_n^{1/4}, \breve{\delta}_n^{1/4}\} \le 0.077$, then

2.1. $\left|I\!\!P\left(\|\boldsymbol{\phi}\| > z\right) - I\!\!P\left(\|\overline{\boldsymbol{\psi}}\| > \overline{z}\right)\right| \le 2.71\delta_n^{1/4} + \delta_z/\sqrt{2} + \delta_\Sigma/2,$

2.2. $\left|I\!\!P\left(\|\boldsymbol{\phi}\| > z\right) - I\!\!P\left(\|\boldsymbol{\psi}\| > \overline{z}\right)\right| \le 2.71\left(\delta_n^{1/4} + \breve{\delta}_n^{1/4}\right) + \delta_z/\sqrt{2} + \delta_\Sigma/2.$

Proof of Theorem B.1. The inequality *1.1* is based on the results of Lemmas C.1, B.4 and B.5:

$$\begin{aligned} I\!\!P\left(\|\boldsymbol{\phi}\| > z\right) &\stackrel{\text{by L. C.1}}{\le} I\!\!P\left(\|\overline{\boldsymbol{\phi}}\| > z - \Delta\right) + 16\Delta^{-3}\delta_n \\ &\stackrel{\text{by L. B.5}}{\le} I\!\!P\left(\|\overline{\boldsymbol{\psi}}\| > z - \Delta\right) + 16\Delta^{-3}\delta_n + \delta_\Sigma/2 \\ &\stackrel{\text{by L. B.4}}{\le} I\!\!P\left(\|\overline{\boldsymbol{\psi}}\| > \overline{z}\right) + 16\Delta^{-3}\delta_n + \delta_\Sigma/2 + (\delta_z + \Delta)\overline{z}^{-1}\sqrt{p/2}. \end{aligned}$$

The inequality *1.2* is implied by the triangle inequality and the sum of two bounds: the bound *1.1* for $\left|I\!\!P\left(\|\boldsymbol{\phi}\| > z\right) - I\!\!P\left(\|\overline{\boldsymbol{\psi}}\| > \overline{z}\right)\right|$ and the bound

$$\left|I\!\!P\left(\|\boldsymbol{\psi}\| > \overline{z}\right) - I\!\!P\left(\|\overline{\boldsymbol{\psi}}\| > \overline{z}\right)\right| \le 16\breve{\delta}_n\Delta^{-3} + \Delta\overline{z}^{-1}\sqrt{p/2},$$

which also follows from *1.1* by taking $\boldsymbol{\phi} := \boldsymbol{\psi}$, $z := \overline{z}$. In this case $\Sigma = \breve{\Sigma}$ and $\delta_\Sigma = \delta_z = 0$.

The second part of the statement follows from the first part by minimising the error term $16\delta_n\Delta^{-3} + \Delta/\sqrt{2}$ w.r.t. Δ. □

Remark B.1. The approximation error in the statements of Theorem B.1 includes three terms, each of them is responsible for a step of derivation: Gaussian approximation, Gaussian comparison and anti-concentration. The value δ_Σ bounds the relation between covariance matrices, δ_z corresponds to the difference between quantiles. $\delta_n^{1/4}$ comes from the Gaussian approximation, under certain conditions this is the biggest term in the expressions *2.1, 2.2* (cf. the proof of Theorem 2.1).

Remark B.2. Here we briefly comment how our results can be compared with what is available in the literature. In the case of i.i.d. vectors $\boldsymbol{\phi}_i$ and $\operatorname{Var}\boldsymbol{\phi}_i \equiv \boldsymbol{I}_p$ Bentkus (2003) obtained the rate $\mathbb{E}\|\boldsymbol{\phi}_1\|^3/\sqrt{n}$ for the error of approximation $\sup_{A\in\mathcal{A}}\bigl|\mathbb{P}(\boldsymbol{\phi}\in A) - \mathbb{P}(\overline{\boldsymbol{\phi}}\in A)\bigr|$, where $\mathcal{A}$ is a class of all Euclidean balls in $\mathbb{R}^p$. Götze (1991) showed for independent vectors $\boldsymbol{\phi}_i$ and their standardized sum $\boldsymbol{\phi}$:

$$\delta_{GAR} \le \begin{cases} C_1\sqrt{p}\sum_{i=1}^n \mathbb{E}\|\boldsymbol{\phi}_i\|^3/\sqrt{n}, & p\in[2,5], \\ C_2 p\sum_{i=1}^n \mathbb{E}\|\boldsymbol{\phi}_i\|^3/\sqrt{n}, & p\ge 6, \end{cases} \tag{B.6}$$

where $\delta_{GAR} \stackrel{\text{def}}{=} \sup_{B\in\mathcal{B}}\bigl|\mathbb{P}(\boldsymbol{\phi}\in B) - \mathbb{P}(\overline{\boldsymbol{\phi}}\in B)\bigr|$ and $\mathcal{B}$ is a class of all measurable convex sets in $\mathbb{R}^p$, the constants $C_1, C_2 > 150$. Bhattacharya and Holmes (2010) argued that the results by Götze (1991) might require more thorough derivation, they obtained the rate $p^{5/2}\sum_{i=1}^n \mathbb{E}\|\boldsymbol{\phi}_i\|^3$ for the previous bound (and $p^{5/2}\mathbb{E}\|\boldsymbol{\phi}_1\|^3/n^{1/2}$ in the i.i.d. case). Chen and Fang (2011) prove that $\delta_{GAR} \le 115\sqrt{p}\sum_{i=1}^n \mathbb{E}\|\boldsymbol{\phi}_i\|^3$ for independent vectors $\boldsymbol{\phi}_i$ with a standardized sum. Götze and Zaitsev (2014) obtained the rate $\mathbb{E}\|\boldsymbol{\phi}_1\|^4/n$ for the Kolmogorov-Smirnov distance between the distributions of $\|\boldsymbol{\phi}\|^2$ and $\|\overline{\boldsymbol{\phi}}\|^2$ for the case of i.i.d. vectors $\boldsymbol{\phi}_i$ with a standardized sum, and for $p \ge 5$ or $p = \infty$. See also Prokhorov and Ulyanov (2013) for the review of the results about normal approximation of quadratic forms in Hilbert space.

Our results ensure the error of the Gaussian approximation of order $2.71\delta_n^{1/4} \le 2.28\bigl\{\sum_{i=1}^n \mathbb{E}\left(\|\boldsymbol{\phi}_i\|^3 + \|\overline{\boldsymbol{\phi}}_i\|^3\right)\bigr\}^{1/4}$. The technique used here is much simpler than in the previous works, and the obtained bounding terms are explicit and only use independence of the $\boldsymbol{\phi}_i$ and $\boldsymbol{\psi}_i$. However, for some special cases, the use of more advanced results about Gaussian approximation may lead to sharper bounds. For instance, for an i.i.d. sample, the GAR error rate $\delta_{GAR} = \sqrt{p^3/n}$ by Bentkus (2003) is better then ours $(p^3/n)^{1/8}$, and in the one-dimensional case Berry-Esseen's theorem would also work better (see Section B.1). In those cases one can improve the overall error bound of the bootstrap approximation by putting δ_{GAR} in place of the sum $16\delta_n\Delta^{-3} + \Delta/\sqrt{2}$. Section B.3 comments how our results can be used to obtain the error rate $\sqrt{p^3/n}$ by using a smoothed quantile function.

B.1 The case of $p = 1$ using Berry-Esseen theorem

Let us consider how the results of Theorem B.1 can be refined in the case $p = 1$ using Berry-Esseen theorem. Introduce similarly to δ_n and $\breve{\delta}_n$ from (B.4) the bounded values

$$\delta_{n,\text{B. E.}} \stackrel{\text{def}}{=} \sum\nolimits_{i=1}^{n} I\!E|\boldsymbol{\phi}_i|^3, \qquad \breve{\delta}_{n,\text{B. E.}} \stackrel{\text{def}}{=} \sum\nolimits_{i=1}^{n} I\!E|\boldsymbol{\psi}_i|^3. \tag{B.7}$$

Due to Berry-Esseen theorem by Berry (1941) and Esseen (1942) it holds

$$\sup_{z \in I\!R} \left| I\!P\left(|\boldsymbol{\phi}| > z\right) - I\!P\left(|\overline{\boldsymbol{\phi}}| > z\right) \right| \leq 2C_0 \frac{\delta_{n,\text{B. E.}}}{(\operatorname{Var} \boldsymbol{\phi})^{3/2}}, \tag{B.8}$$

$$\sup_{z \in I\!R} \left| I\!P\left(|\boldsymbol{\psi}| > z\right) - I\!P\left(|\overline{\boldsymbol{\psi}}| > z\right) \right| \leq 2C_0 \frac{\breve{\delta}_{n,\text{B. E.}}}{(\operatorname{Var} \boldsymbol{\psi})^{3/2}}, \tag{B.9}$$

for the constant $C_0 \in [0.4097, 0.560]$ by Esseen (1956) and Shevtsova (2010).

Lemma B.1. *Under the conditions of Theorem B.1 it holds for* $\overline{z} \geq 1$

1. $\left| I\!P\left(|\boldsymbol{\phi}| > z\right) - I\!P\left(|\overline{\boldsymbol{\psi}}| > \overline{z}\right) \right| \leq 2C_0 \frac{\delta_{n,\text{B. E.}}}{(\operatorname{Var} \boldsymbol{\phi})^{3/2}} + \frac{\delta_\Sigma}{2} + \frac{\delta_z}{\sqrt{2}},$
2. $\left| I\!P\left(|\boldsymbol{\phi}| > z\right) - I\!P\left(|\boldsymbol{\psi}| > \overline{z}\right) \right|$
$$\leq 2C_0 \Big\{ \frac{\delta_{n,\text{B. E.}}}{(\operatorname{Var} \boldsymbol{\phi})^{3/2}} + \frac{\breve{\delta}_{n,\text{B. E.}}}{(\operatorname{Var} \boldsymbol{\psi})^{3/2}} \Big\} + \frac{\delta_\Sigma}{2} + \frac{\delta_z}{\sqrt{2}}. \tag{B.10}$$

Proof of Lemma B.1. Similarly to the proof of Theorem B.1:

$$\begin{aligned} I\!P\left(|\boldsymbol{\phi}| > z\right) &\stackrel{\text{by (B.8)}}{\leq} I\!P\left(|\overline{\boldsymbol{\phi}}| > z\right) + 2C_0 (\operatorname{Var} \boldsymbol{\phi})^{-3/2} \delta_{n,\text{B. E.}} \\ &\stackrel{\text{by L. B.5}}{\leq} I\!P\left(|\overline{\boldsymbol{\psi}}| > z\right) + 2C_0 (\operatorname{Var} \boldsymbol{\phi})^{-3/2} \delta_{n,\text{B. E.}} + \delta_\Sigma/2 \\ &\stackrel{\text{by L. B.4}}{\leq} I\!P\left(|\overline{\boldsymbol{\psi}}| > \overline{z}\right) + 2C_0 (\operatorname{Var} \boldsymbol{\phi})^{-3/2} \delta_{n,\text{B. E.}} + \delta_\Sigma/2 + \delta_z \overline{z}^{-1} 2^{-1/2}. \end{aligned}$$

The analogous chain in the inverse direction finishes the proof of the first part of the statement. The second part is implied by the triangle inequality applied to the first part and again to it with $\boldsymbol{\phi} := \boldsymbol{\psi}$ and $z := \overline{z}$. □

B.2 Gaussian approximation of ℓ_2-norm of a sum of independent vectors

Lemma B.2 (GAR with equal covariance matrices)**.** *For the random vectors* $\boldsymbol{\phi}$ *and* $\overline{\boldsymbol{\phi}}$ *defined in* (B.1), (B.3), *s.t.* $\operatorname{Var} \boldsymbol{\phi} = \operatorname{Var} \overline{\boldsymbol{\phi}}$, *and for* δ_n *given in* (B.4), *it holds for all* $z \geq 2$ *and* $\Delta \in (0, 0.22]$ *:*

$$\begin{aligned} I\!P\left(\|\boldsymbol{\phi}\| > z\right) &\leq I\!P\left(\|\overline{\boldsymbol{\phi}}\| > z - \Delta\right) + 16\Delta^{-3}\delta_n, \\ I\!P\left(\|\boldsymbol{\phi}\| > z\right) &\geq I\!P\left(\|\overline{\boldsymbol{\phi}}\| > z + \Delta\right) - 16\Delta^{-3}\delta_n. \end{aligned}$$

Proof of Lemma C.1. It holds for $z \in I\!R$ $I\!P(\|\boldsymbol{\phi}\| > z) = I\!E\, 1\!\!1\{\|\boldsymbol{\phi}\| > z\}$. The main idea of the proof is to approximate the discontinuous function $1\!\!1\{\|\boldsymbol{\phi}\| > z\}$ by a smooth function $f_{\Delta}(\boldsymbol{\phi}, z)$ and then to apply the Lindeberg's telescopic sum device. Let us introduce a non-negative three times differentiable function $g(\cdot)$, which grows monotonously from 0 to 1:

$$g(x) \stackrel{\text{def}}{=} \begin{cases} 0, & x \le 0, \\ 16x^3/3, & x \in [0, 1/4], \\ 0.5 + 2(x - 0.5) - 16(x - 0.5)^3/3, & x \in [1/4, 3/4], \\ 1 + 16(x-1)^3/3, & x \in [3/4, 1], \\ 1, & x \ge 1. \end{cases} \tag{B.11}$$

It holds for all $x \in I\!R$ $1\!\!1\{x > 1\} \le g(x) \le 1\!\!1\{x > 0\}$. Hence, for the function $f_{\Delta}(\boldsymbol{\phi}, z) \stackrel{\text{def}}{=} g\left((\|\boldsymbol{\phi}\|^2 - z^2)/(2z\Delta)\right)$ with $z, \Delta > 0$, it holds due to $1\!\!1\{\|\boldsymbol{\phi}\| > z\} = 1\!\!1\left\{\left(\|\boldsymbol{\phi}\|^2 - z^2\right)/2 > 0\right\}$:

$$1\!\!1\{\|\boldsymbol{\phi}\| > z + \Delta\} \le 1\!\!1\left\{\|\boldsymbol{\phi}\|^2 > z^2 + 2\Delta z\right\} \le f_{\Delta}(\boldsymbol{\phi}, z) \le 1\!\!1\{\|\boldsymbol{\phi}\| > z\}. \tag{B.12}$$

Due to Lemma B.3 one can apply the Lindeberg's telescopic sum device (see Lindeberg (1922)) in order to approximate $I\!E f_{\Delta}(\boldsymbol{\phi}, z)$ with $I\!E f_{\Delta}(\overline{\boldsymbol{\phi}}, z)$. Define for $k = 2, \ldots, n-1$ the following random sums

$$S_k \stackrel{\text{def}}{=} \sum_{i=1}^{k-1} \overline{\boldsymbol{\phi}}_i + \sum_{i=k+1}^{n} \boldsymbol{\phi}_i, \qquad S_1 \stackrel{\text{def}}{=} \sum_{i=2}^{n} \boldsymbol{\phi}_i, \qquad S_n \stackrel{\text{def}}{=} \sum_{i=1}^{n-1} \overline{\boldsymbol{\phi}}_i.$$

The difference $f_{\Delta}(\boldsymbol{\phi}, z) - f_{\Delta}(\overline{\boldsymbol{\phi}}, z)$ can be represented as the telescopic sum:

$$f_{\Delta}(\boldsymbol{\phi}, z) - f_{\Delta}(\overline{\boldsymbol{\phi}}, z) = \sum\nolimits_{k=1}^{n} \left\{ f_{\Delta}(S_k + \boldsymbol{\phi}_k, z) - f_{\Delta}(S_k + \overline{\boldsymbol{\phi}}_k, z) \right\}.$$

Due to Lemma B.3 and the third order Taylor expansions of $f_{\Delta}(S_k + \boldsymbol{\phi}_k, z)$ and $f_{\Delta}(S_k + \overline{\boldsymbol{\phi}}_k, z)$ w.r.t. the first argument at S_k, it holds for each $k = 1, \ldots, n$:

$$\begin{aligned} \Big| & f_{\Delta}(S_k + \boldsymbol{\phi}_k, z) - f_{\Delta}(S_k + \overline{\boldsymbol{\phi}}_k, z) - \nabla_{\boldsymbol{\phi}} f_{\Delta}(S_k, z)^{\top} (\boldsymbol{\phi}_k - \overline{\boldsymbol{\phi}}_k) \\ & - \frac{1}{2} (\boldsymbol{\phi}_k - \overline{\boldsymbol{\phi}}_k)^{\top} \nabla^2_{\boldsymbol{\phi}} f_{\Delta}(S_k, z) (\boldsymbol{\phi}_k + \overline{\boldsymbol{\phi}}_k) \Big| \le \mathtt{C}(\Delta, z) \left(\|\boldsymbol{\phi}_k\|^3 + \|\overline{\boldsymbol{\phi}}_k\|^3 \right)/6, \end{aligned}$$

where the value $\mathtt{C}(\Delta, z)$ is defined in (B.15). As S_k and $\boldsymbol{\phi}_k - \overline{\boldsymbol{\phi}}_k$ are independent, $I\!E \boldsymbol{\phi}_k = I\!E \overline{\boldsymbol{\phi}}_k = 0$ and $\operatorname{Var} \boldsymbol{\phi}_k = \operatorname{Var} \overline{\boldsymbol{\phi}}_k$, we derive

$$\begin{aligned} \left| I\!E f_{\Delta}(\boldsymbol{\phi}, z) - I\!E f_{\Delta}(\overline{\boldsymbol{\phi}}, z) \right| &= \left| \sum\nolimits_{k=1}^{n} \left\{ I\!E f_{\Delta}(S_k + \boldsymbol{\phi}_k, z) - I\!E f_{\Delta}(S_k + \overline{\boldsymbol{\phi}}_k, z) \right\} \right| \\ &\le \mathtt{C}(\Delta, z) \sum\nolimits_{k=1}^{n} I\!E \left(\|\boldsymbol{\phi}_k\|^3 + \|\overline{\boldsymbol{\phi}}_k\|^3 \right)/6 \\ \text{(by Def. (B.4))} &= \mathtt{C}(\Delta, z) \delta_n / 3. \end{aligned} \tag{B.13}$$

Combining the derived bounds, we obtain:

$$I\!P\left(\|\boldsymbol{\phi}\| \geq z + \Delta\right) \overset{\text{by (B.12)}}{\leq} I\!E f_\Delta(\boldsymbol{\phi}, z) \overset{\text{by (C.15)}}{\leq} I\!E f_\Delta(\overline{\boldsymbol{\phi}}, z) + \frac{\mathtt{C}(\Delta, z)}{3}\delta_n \overset{\text{by (B.12)}}{\leq} I\!P\left(\|\overline{\boldsymbol{\phi}}\| \geq z\right) + \frac{\mathtt{C}(\Delta, z)}{3}\delta_n,$$

or $I\!P\left(\|\boldsymbol{\phi}\| > z\right) \leq I\!P\left(\|\overline{\boldsymbol{\phi}}\| > z - \Delta\right) + \mathtt{C}(\Delta, z-\Delta)\delta_n/3$. Interchanging the arguments $\boldsymbol{\phi}$ and $\overline{\boldsymbol{\phi}}$ implies the inequality in the inverse direction:

$$I\!P\left(\|\boldsymbol{\phi}\| > z\right) \geq I\!P\left(\|\overline{\boldsymbol{\phi}}\| > z + \Delta\right) - \mathtt{C}(\Delta, z)\delta_n/3.$$

Let us bound the constants $\mathtt{C}(\Delta, z)$ and $\mathtt{C}(\Delta, z-\Delta)$ for the function $g(x)$ given above in (B.11). $|g''(x)| \leq 8$ and $|g'''(x)| \leq 32$ for all $x \in I\!R$. By definition (B.15) it holds for $0 < \Delta \leq 0.22$ and $z \geq 2$:

$$\mathtt{C}(\Delta, z) \leq \mathtt{C}(\Delta, z-\Delta) \leq \Delta^{-3} 48.$$

□

Lemma B.3 (A property of the smooth approximant of the indicator)**.** *Let a function $g(x) : I\!R \mapsto I\!R$ be non-negative, monotonously increasing from 0 to 1 and three times differentiable s.t. $g(x) = 0$ for $x < 0$, $g(x) = 1$ for $x \geq 1$. It holds for all $\boldsymbol{\phi}, \boldsymbol{\phi}_0 \in I\!R^p$, $z, \Delta > 0$, for the Euclidean norm $\|\cdot\|$ and for the function*

$$f_\Delta(\boldsymbol{\phi}, z) \overset{\text{def}}{=} g\left(\frac{1}{2z\Delta}(\|\boldsymbol{\phi}\|^2 - z^2)\right) \tag{B.14}$$

$$\left|f_\Delta(\boldsymbol{\phi}_0 + \boldsymbol{\phi}, z) - f_\Delta(\boldsymbol{\phi}_0, z) - \boldsymbol{\phi}^\top \nabla_{\boldsymbol{\phi}} f_\Delta(\boldsymbol{\phi}_0, z) - \boldsymbol{\phi}^\top \nabla^2_{\boldsymbol{\phi}} f_\Delta(\boldsymbol{\phi}_0, z)\boldsymbol{\phi}/2\right| \leq \mathtt{C}(\Delta, z)\|\boldsymbol{\phi}\|^3/3!,$$

where

$$\mathtt{C}(\Delta, z) \overset{\text{def}}{=} \frac{1}{\Delta^3}\left(1 + 2\frac{\Delta}{z}\right)^{1/2}\left\{\left(1 + 2\frac{\Delta}{z}\right)\|g'''\|_\infty + 3\frac{\Delta}{z}\|g''\|_\infty\right\}. \tag{B.15}$$

Proof of Lemma B.3. By the Taylor's formula:

$$f_\Delta(\boldsymbol{\phi}_0 + \boldsymbol{\phi}, z) = f_\Delta(\boldsymbol{\phi}_0, z) + \boldsymbol{\phi}^\top \nabla_{\boldsymbol{\phi}} f_\Delta(\boldsymbol{\phi}_0, z) + \boldsymbol{\phi}^\top \nabla^2_{\boldsymbol{\phi}} f_\Delta(\boldsymbol{\phi}_0, z)\boldsymbol{\phi}/2 + R_3,$$

where R_3 is the 3-d order remainder term. Consider for $\boldsymbol{\gamma} \in I\!R^p : \|\boldsymbol{\gamma}\| = 1$ and $t \in I\!R$ the function $f_\Delta(\boldsymbol{\phi}_0 + t\boldsymbol{\gamma}, z) = g\left(\frac{1}{2z\Delta}(\|\boldsymbol{\phi}_0 + t\boldsymbol{\gamma}\|^2 - z^2)\right)$. It holds

$$|R_3| \leq \frac{\|\boldsymbol{\phi}\|^3}{3!} \sup_{\boldsymbol{\gamma} \in I\!R^p,\, \|\boldsymbol{\gamma}\|=1} \sup_{t \in I\!R} \left|\frac{d^3 f_\Delta(\boldsymbol{\phi}_0 + t\boldsymbol{\gamma}, z)}{dt^3}\right|.$$

Now let us bound the third derivative $\frac{d^3}{dt^3} f_\Delta(\boldsymbol{\phi} + t\boldsymbol{\gamma}, z)$:

$$\begin{aligned}
\frac{df_\Delta(\boldsymbol{\phi} + t\boldsymbol{\gamma}, z)}{dt} &= \frac{\boldsymbol{\gamma}^\top(\boldsymbol{\phi} + t\boldsymbol{\gamma})}{z\Delta} g'\left(\frac{1}{2z\Delta}(\|\boldsymbol{\phi} + t\boldsymbol{\gamma}\|^2 - z^2)\right), \\
\frac{d^2 f_\Delta(\boldsymbol{\phi} + t\boldsymbol{\gamma}, z)}{dt^2} &= \frac{\{\boldsymbol{\gamma}^\top(\boldsymbol{\phi} + t\boldsymbol{\gamma})\}^2}{(z\Delta)^2} g''\left(\frac{1}{2z\Delta}(\|\boldsymbol{\phi} + t\boldsymbol{\gamma}\|^2 - z^2)\right) \\
&\quad + \frac{1}{z\Delta} g'\left(\frac{1}{2z\Delta}(\|\boldsymbol{\phi} + t\boldsymbol{\gamma}\|^2 - z^2)\right), \\
\frac{d^3 f_\Delta(\boldsymbol{\phi} + t\boldsymbol{\gamma}, z)}{dt^3} &= \frac{\{\boldsymbol{\gamma}^\top(\boldsymbol{\phi} + t\boldsymbol{\gamma})\}^3}{(z\Delta)^3} g'''\left(\frac{1}{2z\Delta}(\|\boldsymbol{\phi} + t\boldsymbol{\gamma}\|^2 - z^2)\right) \\
&\quad + 3\frac{\boldsymbol{\gamma}^\top(\boldsymbol{\phi} + t\boldsymbol{\gamma})}{(z\Delta)^2} g''\left(\frac{1}{2z\Delta}(\|\boldsymbol{\phi} + t\boldsymbol{\gamma}\|^2 - z^2)\right).
\end{aligned}$$

Now we use that $g''(x)$ and $g'''(x)$ vanish if $x < 0$ or $x \geq 1$. The inequality $\frac{1}{2z\Delta}(\|\boldsymbol{\phi} + t\boldsymbol{\gamma}\|^2 - z^2) \leq 1$ implies in view of $\|\boldsymbol{\gamma}\| = 1$ that

$$\boldsymbol{\gamma}^\top(\boldsymbol{\phi} + t\boldsymbol{\gamma}) \leq \|\boldsymbol{\phi} + t\boldsymbol{\gamma}\| \leq (2z\Delta + z^2)^{1/2}.$$

Therefore

$$\left|\frac{d^3 f_\Delta(\boldsymbol{\phi}_0 + t\boldsymbol{\gamma}, z)}{dt^3}\right| \leq \frac{1}{\Delta^3}\left(1 + 2\frac{\Delta}{z}\right)^{1/2}\left\{\left(1 + 2\frac{\Delta}{z}\right)\|g'''\|_\infty + 3\frac{\Delta}{z}\|g''\|_\infty\right\}.$$

□

B.3 Results for the smoothed indicator function

Theorem B.2 (Theorem B.1 for a smoothed indicator function)**.** *Under the conditions of Theorem B.1 it holds for all* $\delta_z \in [0, 1]$ *and the function* $f_\Delta(\boldsymbol{\phi}, z)$ *defined in* (B.14)*:*

1. $$\begin{aligned}\left|I\!\!E f_\Delta(\boldsymbol{\phi}, z) - I\!\!E f_\Delta(\overline{\boldsymbol{\psi}}, \overline{z})\right| &\leq \frac{16}{\Delta^3}\delta_n + 2\sqrt{p}\frac{\delta_z}{z} + \sqrt{p}\frac{\delta_z^2}{z^2} + \delta_\Sigma \\ &\leq \frac{16}{\Delta^3}\delta_n + \sqrt{5}\delta_z + \delta_\Sigma \quad \textit{for } z \geq \sqrt{p}.\end{aligned}$$
2. $$\begin{aligned}|I\!\!E f_\Delta(\boldsymbol{\phi}, z) - I\!\!E f_\Delta(\boldsymbol{\psi}, \overline{z})| &\leq \frac{16}{\Delta^3}\left(\delta_n + \breve{\delta}_n\right) + 2\sqrt{p}\frac{\delta_z}{z} + \sqrt{p}\frac{\delta_z^2}{z^2} + \delta_\Sigma \\ &\leq \frac{16}{\Delta^3}\left(\delta_n + \breve{\delta}_n\right) + \sqrt{5}\delta_z + \delta_\Sigma \quad \textit{for } z \geq \sqrt{p}.\end{aligned}$$

Remark B.3. The approximating bounds above do not contain the term proportional to Δ unlike the bound in Theorem B.1. This yields the smaller error terms for the case of the smoothed indicator.

Proof of Theorem B.2. The following inequality is proved in Lemma C.1 (see the expression (C.15)): $\bigl|I\!\!E f_\Delta(\boldsymbol{\phi}, z) - I\!\!E f_\Delta(\overline{\boldsymbol{\phi}}, z)\bigr| \le \mathtt{C}(\Delta, z)\delta_n/3$.

The function $f_\Delta(\boldsymbol{\phi}, z)$ is non-increasing in z:

$$\frac{df_\Delta(\boldsymbol{\phi}, z)}{dz} = -\frac{1}{2\Delta}\left(1 + \frac{\|\boldsymbol{\phi}\|^2}{z^2}\right) g'\left(\frac{1}{2\Delta z}\left(\|\boldsymbol{\phi}\|^2 - z^2\right)\right) \le 0.$$

The definition of $f_\Delta(\boldsymbol{\phi}, z)$ yields for $\overline{z} \ge z$, $a \stackrel{\text{def}}{=} \overline{z}/z \ge 1$ and any $\boldsymbol{\phi}$

$$\begin{aligned} f_\Delta(\boldsymbol{\phi}, \overline{z}) &\le f_\Delta(\boldsymbol{\phi}, z) \le f_\Delta(a\boldsymbol{\phi}, \overline{z}), \\ 0 \le f_\Delta(\boldsymbol{\phi}, z) - f_\Delta(\boldsymbol{\phi}, \overline{z}) &\le f_\Delta(a\boldsymbol{\phi}, \overline{z}) - f_\Delta(\boldsymbol{\phi}, \overline{z}). \end{aligned} \tag{B.16}$$

Lemma B.6 yields for $\delta_z \le z(\sqrt{3/2} - 1)$:

$$\begin{aligned} \bigl|I\!\!E f_\Delta(a\overline{\boldsymbol{\phi}}, \overline{z}) - I\!\!E f_\Delta(\overline{\boldsymbol{\phi}}, \overline{z})\bigr| &\le \sqrt{p}\bigl(\frac{\overline{z}^2}{z^2} - 1\bigr) \le 2\sqrt{p}\frac{\delta_z}{z} + \sqrt{p}\frac{\delta_z^2}{z^2} \\ &\le (1 + \sqrt{3/2})\delta_z \le \sqrt{5}\delta_z \quad \text{for } z \ge \sqrt{p}. \end{aligned}$$

Inequalities similar to (B.16) hold for $\overline{z} \le z$ and $a \stackrel{\text{def}}{=} z/\overline{z}$, therefore, by triangle inequality, bound (B.2) on $\mathtt{C}(\Delta, z)$ and Lemma B.6:

$$\begin{aligned} \bigl|I\!\!E f_\Delta(\boldsymbol{\phi}, z) - I\!\!E f_\Delta(\overline{\boldsymbol{\psi}}, \overline{z})\bigr| &\le \frac{16}{\Delta^3}\delta_n + 2\sqrt{p}\frac{\delta_z}{z} + \sqrt{p}\frac{\delta_z^2}{z^2} + \delta_\Sigma \\ &\le \frac{16}{\Delta^3}\delta_n + \sqrt{5}\delta_z + \delta_\Sigma \quad \text{for } z \ge \sqrt{p}. \end{aligned}$$

The second part of the statement follows from triangle inequality applied to the first inequality and again to the same one with $\boldsymbol{\phi} := \boldsymbol{\psi}$ and $z := \overline{z}$. □

B.4 Gaussian anti-concentration and comparison by the Pinsker's inequality

Lemma B.4 (Anti-concentration bound for ℓ_2 norm of a Gaussian vector)**.** *Let $\overline{\boldsymbol{\phi}} \sim \mathcal{N}(0, \Sigma)$, $\overline{\boldsymbol{\phi}} \in I\!R^p$, then it holds for all $z > 0$ and $0 \le \Delta \le z$:*

$$\begin{aligned} \bigl|I\!\!P\left(\|\overline{\boldsymbol{\phi}}\| \ge z + \Delta\right) - I\!\!P\left(\|\overline{\boldsymbol{\phi}}\| \ge z\right)\bigr| &\le \Delta\sqrt{p}/(z\sqrt{2}) \\ &\le \Delta/\sqrt{2} \quad \textit{for } z \ge \sqrt{p}. \end{aligned}$$

Proof of Lemma B.4. It holds $I\!\!P\left(\|\overline{\boldsymbol{\phi}}\| \ge z + \Delta\right) = I\!\!P\left(\|\overline{\boldsymbol{\phi}}_\Delta\| \ge z\right)$, where $\overline{\boldsymbol{\phi}}_\Delta \stackrel{\text{def}}{=} \overline{\boldsymbol{\phi}}_{\frac{z}{z+\Delta}}$. The Kullback-Leibler divergence between $I\!\!P_1 \stackrel{\text{def}}{=} \mathcal{N}(0, \Sigma)$ and $I\!\!P_2 \stackrel{\text{def}}{=} \mathcal{N}\bigl(0, \Sigma\frac{z^2}{(z+\Delta)^2}\bigr)$ is equal to

$$\begin{aligned} \mathrm{KL}(I\!\!P_1, I\!\!P_2) &= p\left\{(\Delta/z)^2 + 2(\Delta/z) - 2\log(1 + \Delta/z)\right\}/2 \\ &\le p(\Delta/z)^2 \quad \text{for } 0 \le \Delta \le z. \end{aligned}$$

We use the Pinsker's inequality in the following form (see, e.g., Tsybakov (2009), pp. 88, 132): for a measurable space $(\Omega, \mathcal{F})$ and two measures on it $I\!P_1, I\!P_2$:

$$\sup_{A \in \mathcal{F}} |I\!P_1(A) - I\!P_2(A)| \leq \sqrt{\mathrm{KL}(I\!P_1, I\!P_2)/2}. \tag{B.17}$$

Therefore, it holds:

$$\left| I\!P\left(\|\overline{\boldsymbol{\phi}}\| \geq z + \Delta\right) - I\!P\left(\|\overline{\boldsymbol{\phi}}\| \geq z\right)\right| \leq \sqrt{\mathrm{KL}(I\!P_1, I\!P_2)/2} \leq \Delta\sqrt{p}/(z\sqrt{2}).$$

□

Lemma B.5 (Comparison of the Euclidian norms of Gaussian vectors)**.** *Let* $\overline{\boldsymbol{\psi}}_1 \sim \mathcal{N}(0, \Sigma_1)$ *and* $\overline{\boldsymbol{\psi}}_2 \sim \mathcal{N}(0, \Sigma_2)$ *belong to* $I\!R^p$*, and*

$$\left\|\Sigma_2^{-1/2}\Sigma_1\Sigma_2^{-1/2} - \boldsymbol{I}_p\right\| \leq 1/2, \quad \textit{and} \quad \mathrm{tr}\left\{\left(\Sigma_2^{-1/2}\Sigma_1\Sigma_2^{-1/2} - \boldsymbol{I}_p\right)^2\right\} \leq \rho_\Sigma^2, \tag{B.18}$$

for some $\rho_\Sigma^2 \geq 0$*. Then it holds*

$$\sup_{z \in R}\left| I\!P\left(\|\overline{\boldsymbol{\psi}}_1\| \geq z\right) - I\!P\left(\|\overline{\boldsymbol{\psi}}_2\| \geq z\right)\right| \leq \rho_\Sigma/2.$$

Proof of Lemma B.5. Let $I\!P_1 = \mathcal{N}(0, \Sigma_1)$ and $I\!P_2 = \mathcal{N}(0, \Sigma_2)$. Denote $G \stackrel{\mathrm{def}}{=} \Sigma_2^{-1/2}\Sigma_1\Sigma_2^{-1/2}$, then the Kullback-Leibler divergence between $I\!P_1$ and $I\!P_2$ is equal to

$$\begin{aligned}\mathrm{KL}(I\!P_1, I\!P_2) &= -0.5\log\{\det(G)\} + 0.5\,\mathrm{tr}\{G - \boldsymbol{I}_p\} \\ &= 0.5\sum\nolimits_{j=1}^{p}\{\lambda_j - \log(\lambda_j + 1)\},\end{aligned}$$

where $\lambda_p \leq \cdots \leq \lambda_1$ are the eigenvalues the matrix $G - \boldsymbol{I}_p$. By conditions of the lemma $|\lambda_1| \leq 1/2$, and it holds:

$$\mathrm{KL}(I\!P_1, I\!P_2) \leq 0.5\sum\nolimits_{j=1}^{p}\lambda_j^2 = 0.5\,\mathrm{tr}\{(G - \boldsymbol{I}_p)^2\} \leq \rho_\Sigma^2/2, \tag{B.19}$$

which finishes the proof due to the Pinsker's inequality (B.17). □

Remark B.4. Barsov and Ul'yanov (1987) obtained estimates for the difference of two normal measures of Euclidean balls in a real separable Hilbert space. These results applied to the setting of Lemma B.5 lead to a similar approximation bound as the one in Lemma B.5.

Lemma B.6 (Gaussian comparison, smoothed version)**.** *Let* $\overline{\boldsymbol{\psi}}_1 \sim \mathcal{N}(0, \Sigma_1)$ *and* $\overline{\boldsymbol{\psi}}_2 \sim \mathcal{N}(0, \Sigma_2)$ *belong to* $I\!R^p$*, and for some* $\rho_\Sigma^2 \geq 0$*:*

$$\left\|\Sigma_2^{-1/2}\Sigma_1\Sigma_2^{-1/2} - \boldsymbol{I}_p\right\| \leq 1/2, \quad \textit{and} \quad \mathrm{tr}\left\{\left(\Sigma_2^{-1/2}\Sigma_1\Sigma_2^{-1/2} - \boldsymbol{I}_p\right)^2\right\} \leq \rho_\Sigma^2. \tag{B.20}$$

Then it holds for any function $f(\boldsymbol{x}) : I\!R^p \mapsto I\!R$ *s.t.* $|f(\boldsymbol{x})| \leq 1$*:*

$$\left|I\!E f(\overline{\boldsymbol{\psi}}_1) - I\!E f(\overline{\boldsymbol{\psi}}_2)\right| \leq \rho_\Sigma.$$

Proof of Lemma B.6. Let $I\!P_1 = \mathcal{N}(0, \Sigma_1)$ and $I\!P_2 = \mathcal{N}(0, \Sigma_2)$. Due to $|f(\boldsymbol{x})| \leq 1$ and Pinsker's inequality (B.17) it holds:

$$\begin{aligned}\left|I\!E f(\overline{\boldsymbol{\psi}}_1) - I\!E f(\overline{\boldsymbol{\psi}}_2)\right| &\leq \int_{I\!R^p} |f(\boldsymbol{x})| \cdot |dI\!P_1(\boldsymbol{x}) - dI\!P_2(\boldsymbol{x})| \\ &\leq \int_{I\!R^p} |dI\!P_1(\boldsymbol{x}) - dI\!P_2(\boldsymbol{x})| \leq 2\,\sqrt{\mathrm{KL}(I\!P_1, I\!P_2)/2}.\end{aligned}$$

Finally, as in (B.19), $2\sqrt{\mathrm{KL}(I\!P_1, I\!P_2)/2} \leq \rho_{\Sigma}$. □

Appendix C

Approximation of the joint distributions of ℓ_2-norms

In this chapter we obtain an approximation bound between the joint distributions of ℓ_2-norms of two (independent) sets of random vectors: $\{\|\boldsymbol{\phi}_1\|, \ldots, \|\boldsymbol{\phi}_K\|\}$ and $\{\|\boldsymbol{\psi}_1\|, \ldots, \|\boldsymbol{\psi}_K\|\}$, where for each $k = 1, \ldots, K$ $\boldsymbol{\phi}_k, \boldsymbol{\psi}_k \in I\!R^{p_k}$ and equal to sums of independent centered vectors.

Consider K random centered vectors $\boldsymbol{\phi}_k \in I\!R^{p_k}$ for $k = 1, \ldots, K$. Each vector equals to a sum of n centered independent vectors:

$$\begin{aligned} \boldsymbol{\phi}_k &= \boldsymbol{\phi}_{k,1} + \cdots + \boldsymbol{\phi}_{k,n}, \\ I\!E\boldsymbol{\phi}_k &= I\!E\boldsymbol{\phi}_{k,i} = 0 \quad \forall\, 1 \le i \le n. \end{aligned} \tag{C.1}$$

Introduce similarly the vectors $\boldsymbol{\psi}_k \in I\!R^{p_k}$ for $k = 1, \ldots, K$:

$$\begin{aligned} \boldsymbol{\psi}_k &= \boldsymbol{\psi}_{k,1} + \cdots + \boldsymbol{\psi}_{k,n}, \\ I\!E\boldsymbol{\psi}_k &= I\!E\boldsymbol{\psi}_{k,i} = 0 \quad \forall\, 1 \le i \le n, \end{aligned} \tag{C.2}$$

with the same independence properties as $\boldsymbol{\phi}_{k,i}$, and also independent of all $\boldsymbol{\phi}_{k,i}$.

The goal of this chapter is to compare the joint distributions of the ℓ_2-norms of the sets of vectors $\boldsymbol{\phi}_k$ and $\boldsymbol{\psi}_k$, $k = 1, \ldots, K$ (i.e. the probability laws $\mathcal{L}(\|\boldsymbol{\phi}_1\|, \ldots, \|\boldsymbol{\phi}_K\|)$ and $\mathcal{L}(\|\boldsymbol{\psi}_1\|, \ldots, \|\boldsymbol{\psi}_K\|)$), assuming that their correlation structures are close to each other.

Denote

$$
\begin{aligned}
p_{\max} &\stackrel{\text{def}}{=} \max_{1\le k\le K} p_k, \quad p_{\text{sum}} \stackrel{\text{def}}{=} p_1 + \cdots + p_K, \\
\lambda^2_{\phi,\max} &\stackrel{\text{def}}{=} \max_{1\le k\le K} \| \operatorname{Var}(\boldsymbol{\phi}_j)\|, \quad \lambda^2_{\psi,\max} \stackrel{\text{def}}{=} \max_{1\le k\le K} \| \operatorname{Var}(\boldsymbol{\psi}_j)\|, \\
z_{\max} &\stackrel{\text{def}}{=} \max_{1\le k\le K} z_k, \quad z_{\min} \stackrel{\text{def}}{=} \min_{1\le k\le K} z_k, \\
\delta_{z,\max} &\stackrel{\text{def}}{=} \max_{1\le k\le K} \delta_{z_k}, \quad \delta_{z,\min} \stackrel{\text{def}}{=} \min_{1\le k\le K} \delta_{z_k},
\end{aligned}
$$

let also

$$
\begin{aligned}
\Delta_\varepsilon \stackrel{\text{def}}{=} & \left(\frac{p^3_{\max}}{n}\right)^{1/8} \log^{9/16}(K) \log^{3/8}(n p_{\text{sum}}) z^{1/8}_{\min} \\
& \times \max\{\lambda_{\phi,\max}, \lambda_{\psi,\max}\}^{3/4} \log^{-1/8}(5n^{1/2}).
\end{aligned} \tag{C.3}
$$

The following conditions are required for the Proposition C.1

(C1) *For some* $\mathtt{g}_k, \nu_k, \mathfrak{c}_\phi, \mathfrak{c}_\psi > 0$ *and for all* $i = 1, \ldots, n$, $k = 1, \ldots, K$

$$
\begin{aligned}
\sup_{\substack{\boldsymbol{\gamma}_k \in I\!R^{p_k}, \\ \|\boldsymbol{\gamma}_k\|=1}} \log I\!E \exp\left\{\lambda\sqrt{n}\boldsymbol{\gamma}_k^\top \boldsymbol{\phi}_{k,i}/\mathfrak{c}_\phi\right\} &\le \lambda^2\nu_k^2/2, \quad |\lambda| < \mathtt{g}_k, \\
\sup_{\substack{\boldsymbol{\gamma}_k \in I\!R^{p_k}, \\ \|\boldsymbol{\gamma}_k\|=1}} \log I\!E \exp\left\{\lambda\sqrt{n}\boldsymbol{\gamma}_k^\top \boldsymbol{\psi}_{k,i}/\mathfrak{c}_\psi\right\} &\le \lambda^2\nu_k^2/2, \quad |\lambda| < \mathtt{g}_k,
\end{aligned}
$$

where $\mathfrak{c}_\phi \ge \mathtt{C}\lambda_{\phi,\max}$ *and* $\mathfrak{c}_\psi \ge \mathtt{C}\lambda_{\phi,\max}$.

(C2) *For some* $\delta^2_\Sigma \ge 0$

$$
\max_{1\le k_1, k_2\le K} \left\|\operatorname{Cov}(\boldsymbol{\phi}_{k_1}, \boldsymbol{\phi}_{k_2}) - \operatorname{Cov}(\boldsymbol{\psi}_{k_1}, \boldsymbol{\psi}_{k_2})\right\|_{\max} \le \delta^2_\Sigma. \tag{C.4}
$$

Proposition C.1 (Approximation of the joint distributions of ℓ_2-norms)**.** *Consider the centered random vectors* $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_K$ *and* $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_K$ *given in* (C.1), (C.2)*. Let the conditions (C1) and (C2) be fulfilled, and the values* $z_k \ge \sqrt{p_k + \Delta_\varepsilon}$ *and* $\delta_{z_k} \ge 0$ *be s.t.* $\mathtt{C}\max\{n^{-1/2}, \delta_{z,\max}\} \le \Delta_\varepsilon \le \mathtt{C} z^{-1}_{\max}$ *, then it holds with dominating probability*

$$
\begin{aligned}
I\!P\left(\bigcup\nolimits_{k=1}^K \{\|\boldsymbol{\phi}_k\| > z_k\}\right) - I\!P\left(\bigcup\nolimits_{k=1}^K \{\|\boldsymbol{\psi}_k\| > z_k - \delta_{z_k}\}\right) &\ge -\Delta_{\ell_2}, \\
I\!P\left(\bigcup\nolimits_{k=1}^K \{\|\boldsymbol{\phi}_k\| > z_k\}\right) - I\!P\left(\bigcup\nolimits_{k=1}^K \{\|\boldsymbol{\psi}_k\| > z_k + \delta_{z_k}\}\right) &\le \Delta_{\ell_2}
\end{aligned}
$$

for the deterministic non-negative value

$$\begin{aligned}\Delta_{\ell_2} &\le 12.5\mathtt{C}\left(\frac{p_{\max}^3}{n}\right)^{1/8}\log^{9/8}(K)\log^{3/8}(np_{\mathrm{sum}})\max\left\{\lambda_{\phi,\max},\lambda_{\psi,\max}\right\}^{3/4}\\ &+ 3.2\mathtt{C}\delta_\Sigma^2\left(\frac{p_{\max}^3}{n}\right)^{1/4}p_{\max}z_{\min}^{1/2}\log^2(K)\log^{3/4}(np_{\mathrm{sum}})\max\left\{\lambda_{\phi,\max},\lambda_{\psi,\max}\right\}^{7/2}\\ &\le 25\mathtt{C}\left(\frac{p_{\max}^3}{n}\right)^{1/8}\log^{9/8}(K)\log^{3/8}(np_{\mathrm{sum}})\max\left\{\lambda_{\phi,\max},\lambda_{\psi,\max}\right\}^{3/4},\end{aligned}$$

where the last inequality holds for

$$\delta_\Sigma^2 \le 4\mathtt{C}\left(\frac{n}{p_{\max}^{13}}\right)^{1/8}\log^{-7/8}(K)\log^{-3/8}(np_{\mathrm{sum}})\left(\max\left\{\lambda_{\phi,\max},\lambda_{\psi,\max}\right\}\right)^{-11/4}.$$

The proof of this proposition is given in Section C.4.

Remark C.1. The approximating error term Δ_{ℓ_2} consists of three errors, which correspond to: the Gaussian approximation result (Lemma C.1), Gaussian comparison (Lemma C.6), and anti-concentration inequality (Lemma C.7). The bound on Δ_{ℓ_2} above implies that the number K of the random vectors $\boldsymbol{\phi}_1,\ldots,\boldsymbol{\phi}_K$ should satisfy $\log K \ll (n/p_{\max}^3)^{1/12}$ in order to keep the approximating error term Δ_{ℓ_2} small. This condition can be relaxed by using a sharper Gaussian approximation result. For instance, using in Lemma C.1 the Slepian-Stein technique plus induction argument from the recent paper by Chernozhukov et al. (2014b) instead of the Lindeberg's approach, would lead to the improved bound: $\mathtt{C}\left(\frac{p_{\max}^3}{n}\right)^{1/6}$ multiplied by a logarithmic term.

C.1 Joint Gaussian approximation of ℓ_2-norms by Lindeberg's method

Introduce the following random vectors from $\mathbb{R}^{p_{\mathrm{sum}}}$:

$$\begin{aligned}&\boldsymbol{\Phi} \stackrel{\mathrm{def}}{=} \left(\boldsymbol{\phi}_1^\top,\ldots,\boldsymbol{\phi}_K^\top\right)^\top, \quad \boldsymbol{\Phi}_i \stackrel{\mathrm{def}}{=} \left(\boldsymbol{\phi}_{1,i}^\top,\ldots,\boldsymbol{\phi}_{K,i}^\top\right)^\top,\ i=1,\ldots,n,\\ &\boldsymbol{\Phi}=\sum\nolimits_{i=1}^n\boldsymbol{\Phi}_i, \qquad\qquad \mathbb{E}\boldsymbol{\Phi}=\mathbb{E}\boldsymbol{\Phi}_i=0.\end{aligned} \tag{C.5}$$

Define their Gaussian analogs as follows:

$$\overline{\boldsymbol{\Phi}}_i \stackrel{\mathrm{def}}{=} \left(\overline{\boldsymbol{\phi}}_{1,i}^\top,\ldots,\overline{\boldsymbol{\phi}}_{K,i}^\top\right)^\top, \qquad \overline{\boldsymbol{\Phi}} \stackrel{\mathrm{def}}{=} \left(\overline{\boldsymbol{\phi}}_1^\top,\ldots,\overline{\boldsymbol{\phi}}_K^\top\right)^\top=\sum\nolimits_{i=1}^n\overline{\boldsymbol{\Phi}}_i, \tag{C.6}$$

$$\overline{\boldsymbol{\Phi}}_i \sim \mathcal{N}(0,\operatorname{Var}\boldsymbol{\Phi}_i), \qquad \overline{\boldsymbol{\Phi}} \sim \mathcal{N}(0,\operatorname{Var}\boldsymbol{\Phi}), \tag{C.7}$$

$$\overline{\boldsymbol{\phi}}_{k,i} \sim \mathcal{N}(0,\operatorname{Var}\boldsymbol{\phi}_{k,i}), \qquad \overline{\boldsymbol{\phi}}_k \stackrel{\mathrm{def}}{=} \sum\nolimits_{i=1}^n\overline{\boldsymbol{\phi}}_{k,i} \sim \mathcal{N}(0,\operatorname{Var}\boldsymbol{\phi}_k). \tag{C.8}$$

Lemma C.1 (Joint GAR with equal covariance matrices)**.** *Consider the sets of random vectors* $\boldsymbol{\phi}_j$ *and* $\overline{\boldsymbol{\phi}}_j$, $j = 1, \ldots, K$ *defined in* (C.1)*, and* (C.5)– (C.8)*. If the conditions of Lemmas C.3 are C.4 are fulfilled, then it holds for all* $\Delta, \beta > 0$, $z_j \geq \max\left\{\Delta + \sqrt{p_j}, 2.25\log(K)/\beta\right\}$ *with dominating probability*

$$
\begin{aligned}
I\!P\left(\bigcup\nolimits_{j=1}^{K}\left\{\|\boldsymbol{\phi}_j\| > z_j\right\}\right) &\leq I\!P\left(\bigcup\nolimits_{j=1}^{K}\left\{\|\overline{\boldsymbol{\phi}}_j\| > z_j - \Delta - \frac{3\log(K)}{2\beta}\right\}\right) + \delta_{3,\phi}(\Delta,\beta), \\
I\!P\left(\bigcup\nolimits_{j=1}^{K}\left\{\|\boldsymbol{\phi}_j\| > z_j\right\}\right) &\geq I\!P\left(\bigcup\nolimits_{j=1}^{K}\left\{\|\overline{\boldsymbol{\phi}}_j\| > z_j + \Delta + \frac{3\log(K)}{2\beta}\right\}\right) - \delta_{3,\phi}(\Delta,\beta)
\end{aligned}
$$

for $\delta_{3,\phi}(\Delta,\beta) \leq \mathtt{C}\left(\frac{1}{\Delta^3} + \frac{\beta}{\Delta^2} + \frac{\beta^2}{\Delta}\right)\left\{\frac{p_{\max}^3}{n}\log(K)\log^3(np_{\mathrm{sum}})\right\}^{1/2}$ *given in* (C.15)*.*

Proof of Lemma C.1.

$$
I\!P\left(\bigcup\nolimits_{j=1}^{K}\left\{\|\boldsymbol{\phi}_j\| > z_j\right\}\right) = I\!E\, 1\!\!1\big(\max\nolimits_{1\leq j\leq K}\left\{\|\boldsymbol{\phi}_j\|^2 - z_j^2\right\} > 0\big).
$$

Let us approximate the $\max_{1\leq j\leq K}$ function using the smooth maximum:

$$
\begin{aligned}
&h_\beta\left(\{x_j\}\right) \stackrel{\mathrm{def}}{=} \beta^{-1}\log\left(\sum\nolimits_{j=1}^{K}\mathrm{e}^{\beta x_j}\right) \text{ for } \beta > 0,\, x_j \in I\!R, \\
&h_\beta\left(\{x_j\}\right) - \beta^{-1}\log(K) \leq \max_{1\leq j\leq K}\{x_j\} \leq h_\beta\left(\{x_j\}\right). \qquad \text{(C.9)}
\end{aligned}
$$

The indicator function $1\!\!1\{x > 0\}$ is approximated with the three times differentiable function $g(x)$ growing monotonously from 0 to 1:

$$
g(x) \stackrel{\mathrm{def}}{=} \begin{cases} 0, & x \leq 0, \\ 16x^3/3, & x \in [0, 1/4], \\ 0.5 + 2(x-0.5) - 16(x-0.5)^3/3, & x \in [1/4, 3/4], \\ 1 + 16(x-1)^3/3, & x \in [3/4, 1], \\ 1, & x \geq 1. \end{cases}
$$

It holds for all $x \in I\!R$ and $\Delta > 0$

$$
1\!\!1\left\{x > \Delta\right\} \leq g(x/\Delta) \leq 1\!\!1\left\{x/\Delta > 0\right\}.
$$

Therefore

$$\begin{aligned}
&\mathbb{P}\left(\max_{1\le j\le K}\left\{\|\boldsymbol{\phi}_j\| - z_j\right\} > \Delta\right) \\
&\quad\le \mathbb{E}\,\mathbb{1}\left(\max_{1\le j\le K}\left\{\frac{\|\boldsymbol{\phi}_j\|^2 - z_j^2}{2z_j}\right\} > \Delta\right) \\
&\quad\le \mathbb{E}g\left(\max_{1\le j\le K}\left\{\frac{\|\boldsymbol{\phi}_j\|^2 - z_j^2}{2z_j\Delta}\right\}\right) \\
&\quad\le \mathbb{E}g\left(\frac{1}{\Delta\beta}\log\left\{\sum\nolimits_{j=1}^{K}\exp\left[\beta\frac{\|\boldsymbol{\phi}_j\|^2 - z_j^2}{2z_j}\right]\right\}\right) \qquad \text{(C.10)} \\
&\quad\le \mathbb{E}g\left(\max_{1\le j\le K}\left\{\frac{\|\boldsymbol{\phi}_j\|^2 - z_j^2}{2z_j\Delta}\right\} + \frac{\log(K)}{\beta\Delta}\right) \\
&\quad\le \mathbb{E}\,\mathbb{1}\left(\max_{1\le j\le K}\left\{\frac{\|\boldsymbol{\phi}_j\|^2 - z_j^2}{2z_j}\right\} > -\frac{\log(K)}{\beta}\right) \\
&\quad\le \mathbb{P}\left(\max_{1\le j\le K}\left\{\|\boldsymbol{\phi}_j\| - z_j\right\} > -1.5\frac{\log(K)}{\beta}\right), \qquad \text{(C.11)}
\end{aligned}$$

where the last inequality holds for $z_j \ge 2.25\log(K)/\beta$. Denote

$$\boldsymbol{z} \stackrel{\text{def}}{=} (z_1, \ldots, z_K)^\top \in \mathbb{R}^K,\ z_j > 0.$$

Introduce the function $F_{\Delta,\beta}(\Phi, \boldsymbol{z}) : \mathbb{R}^{p_{\text{sum}}} \times \mathbb{R}^K \mapsto \mathbb{R}$:

$$F_{\Delta,\beta}(\Phi, \boldsymbol{z}) \stackrel{\text{def}}{=} g\left(\frac{1}{\Delta\beta}\log\left\{\sum\nolimits_{j=1}^{K}\exp\left[\beta\frac{\|\boldsymbol{\phi}_j\|^2 - z_j^2}{2z_j}\right]\right\}\right) \qquad \text{(C.12)}$$

Then by (C.10) and (C.11)

$$\begin{aligned}
&\mathbb{P}\left(\max_{1\le j\le K}\left\{\|\boldsymbol{\phi}_j\| - z_j\right\} > \Delta\right) \\
&\qquad\le \mathbb{E}F_{\Delta,\beta}(\Phi, \boldsymbol{z}) \qquad \text{(C.13)} \\
&\qquad\le \mathbb{P}\left(\max_{1\le j\le K}\left\{\|\boldsymbol{\phi}_j\| - z_j\right\} > -\frac{3\log(K)}{2\beta}\right). \qquad \text{(C.14)}
\end{aligned}$$

Lemma C.5 checks that $F_{\Delta,\beta}(\cdot, \boldsymbol{z})$ admits applying the Lindeberg's telescopic sum device (see Lindeberg (1922)) in order to approximate $\mathbb{E}F_{\Delta,\beta}(\Phi, \boldsymbol{z})$ with $\mathbb{E}F_{\Delta,\beta}(\overline{\Phi}, \boldsymbol{z})$. Define for $q = 2, \ldots, n-1$ the following $\mathbb{R}^{p_{\text{sum}}}$-valued random sums:

$$S_q \stackrel{\text{def}}{=} \sum_{i=1}^{q-1}\overline{\Phi}_i + \sum_{i=q+1}^{n}\Phi_i, \qquad S_1 \stackrel{\text{def}}{=} \sum_{i=2}^{n}\Phi_i, \qquad S_n \stackrel{\text{def}}{=} \sum_{i=1}^{n-1}\overline{\Phi}_i.$$

The difference $F_{\Delta,\beta}(\Phi, \boldsymbol{z}) - F_{\Delta,\beta}(\overline{\Phi}, \boldsymbol{z})$ can be represented as the telescopic sum:

$$F_{\Delta,\beta}(\Phi, \boldsymbol{z}) - F_{\Delta,\beta}(\overline{\Phi}, \boldsymbol{z}) = \sum\nolimits_{i=1}^{n}\left\{F_{\Delta,\beta}(S_i + \Phi_i, \boldsymbol{z}) - F_{\Delta,\beta}(S_i + \overline{\Phi}_i, \boldsymbol{z})\right\}.$$

The third order Taylor expansions of $F_{\Delta,\beta}(S_i+\Phi_i,\boldsymbol{z})$ and $F_{\Delta,\beta}(S_i+\overline{\Phi}_i,\boldsymbol{z})$ w.r.t. the first argument at S_i, and Lemma C.5 imply for each $i=1,\ldots,n$:

$$\begin{aligned}
&\Big|F_{\Delta,\beta}(S_i+\Phi_i,\boldsymbol{z}) - F_{\Delta,\beta}(S_i+\overline{\Phi}_i,\boldsymbol{z}) - \nabla_{\Phi}F_{\Delta,\beta}(S_i,\boldsymbol{z})^\top(\Phi_i-\overline{\Phi}_i) \\
&\quad - \frac{1}{2}(\Phi_i-\overline{\Phi}_i)^\top\nabla^2_{\Phi}F_{\Delta,\beta}(S_i,\boldsymbol{z})(\Phi_i+\overline{\Phi}_i)\Big| \\
&\le \frac{\mathtt{C}_3(\Delta,\beta)}{6}\left(\max_{1\le j\le K}\left\{\|S_{j,i}+\boldsymbol{\phi}_{j,i}\|^3\right\}\|\Phi_i\|^3_{\max} + \max_{1\le j\le K}\left\{\|S_{j,i}+\overline{\boldsymbol{\phi}}_{j,i}\|^3\right\}\|\overline{\Phi}_i\|^3_{\max}\right),
\end{aligned}$$

where the value $\mathtt{C}_3(\Delta,\beta)$ is defined in Lemma C.5, and the random vectors $S_{j,i}\in\mathbb{R}^{p_j}$ for $j=1,\ldots,K$ are s.t. for all $i=1,\ldots,n$

$$S_i = \left(S_{1,i}^\top, S_{2,i}^\top, \ldots, S_{K,i}^\top\right)^\top.$$

By their construction S_i and $\Phi_i-\overline{\Phi}_i$ are independent, $\mathbb{E}\Phi_i = \mathbb{E}\overline{\Phi}_i = 0$ and $\operatorname{Var}\Phi_i = \operatorname{Var}\overline{\Phi}_i$, therefore

$$\begin{aligned}
&\left|\mathbb{E}F_{\Delta,\beta}(\Phi,\boldsymbol{z}) - \mathbb{E}F_{\Delta,\beta}(\overline{\Phi},\boldsymbol{z})\right| \\
&= \left|\sum\nolimits_{i=1}^n\left\{\mathbb{E}H_{\Delta}(S_i+\Phi_i,\boldsymbol{z}) - \mathbb{E}H_{\Delta}(S_i+\overline{\Phi}_i,\boldsymbol{z})\right\}\right| \\
&\le \frac{\mathtt{C}_3(\Delta,\beta)}{6}\sum_{i=1}^n\mathbb{E}\left(\max_{1\le j\le K}\left\{\|S_{j,i}+\boldsymbol{\phi}_{j,i}\|^3\right\}\|\Phi_i\|^3_{\max} + \max_{1\le j\le K}\left\{\|S_{j,i}+\overline{\boldsymbol{\phi}}_{j,i}\|^3\right\}\|\overline{\Phi}_i\|^3_{\max}\right).
\end{aligned}$$

Lemma C.4 implies for all $i=1,\ldots,n$ with probability $\ge 1-2\mathrm{e}^{-\mathtt{x}}$

$$\left(\mathbb{E}\max_{1\le j\le K}\left\{\|S_{j,i}+\boldsymbol{\phi}_{j,i}\|^6\right\}\right)^{1/2} \le \mathtt{C}\nu_0\max_{1\le j\le K}\|\operatorname{Var}^{1/2}(\boldsymbol{\phi}_j)\|^3\sqrt{p_{\max}\log(K)}(p_{\max}+6\mathtt{x}),$$

and the same bound holds for $\left(\mathbb{E}\max_{1\le j\le K}\left\{\|S_{j,i}+\overline{\boldsymbol{\phi}}_{j,i}\|^6\right\}\right)^{1/2}$. Denote

$$\delta_{\max,\phi} \stackrel{\mathrm{def}}{=} \frac{1}{2}\sum_{i=1}^n\left\{\mathbb{E}\left(\|\Phi_i\|^6_{\max}\right)\right\}^{1/2} + \left\{\mathbb{E}\left(\|\overline{\Phi}_i\|^6_{\max}\right)\right\}^{1/2}.$$

By Lemma C.3 it holds for $t = (\mathtt{x}+\log(p_{\mathrm{sum}}))^3\left(\sqrt{2}\mathfrak{c}_{\phi}\nu_0\right)^6 n^{-3}$ with probability $\ge 1-\mathrm{e}^{-\mathtt{x}}$

$$\|\Phi_i\|^6_{\max}\le t, \quad \|\overline{\Phi}_i\|^6_{\max}\le t.$$

If $\mathtt{x}=\mathtt{C}\log n$, then the last bound on $\left|\mathbb{E}F_{\Delta,\beta}(\Phi,\boldsymbol{z}) - \mathbb{E}F_{\Delta,\beta}(\overline{\Phi},\boldsymbol{z})\right|$ continues with probability $\ge 1-6\exp(-\mathtt{x})$ as follows

$$\begin{aligned}
&\left|\mathbb{E}F_{\Delta,\beta}(\Phi,\boldsymbol{z}) - \mathbb{E}F_{\Delta,\beta}(\overline{\Phi},\boldsymbol{z})\right| \\
&\le \mathtt{C}\frac{\mathtt{C}_3(\Delta,\beta)}{3}\sqrt{p^3_{\max}\log(K)}\delta_{\max,\phi}\max_{1\le j\le K}\|\operatorname{Var}^{1/2}(\boldsymbol{\phi}_j)\|^3 \\
&\le \frac{\mathtt{C}}{3}\left(\frac{1}{\Delta^3}+\frac{\beta}{\Delta^2}+\frac{\beta^2}{\Delta}\right)\frac{p^{3/2}_{\max}}{n^{1/2}}\log^{1/2}(K)\log^{3/2}(np_{\mathrm{sum}})\max_{1\le j\le K}\|\operatorname{Var}^{1/2}(\boldsymbol{\phi}_j)\|^3\left(2\nu_0^2\mathfrak{c}_{\phi}^2\right)^{3/2} \\
&\stackrel{\mathrm{def}}{=} \delta_{3,\phi}(\Delta,\beta).
\end{aligned} \tag{C.15}$$

The derived bounds imply:

$$\begin{aligned}
&I\!P\left(\bigcup\nolimits_{j=1}^{K}\left\{\|\boldsymbol{\phi}_j\| > z_j\right\}\right) \\
&\quad \overset{\text{by (C.13)}}{\leq} I\!E F_{\Delta,\beta}\left(\Phi, \boldsymbol{z} - \Delta\mathbf{1}_K\right) \\
&\quad \overset{\text{by (C.15)}}{\leq} I\!E F_{\Delta,\beta}\left(\overline{\Phi}, \boldsymbol{z} - \Delta\mathbf{1}_K\right) + \delta_{3,\phi}(\Delta,\beta) \qquad \text{(C.16)} \\
&\quad \overset{\text{by (C.14)}}{\leq} I\!P\left(\bigcup\nolimits_{j=1}^{K}\left\{\|\overline{\boldsymbol{\phi}}_j\| > z_j - \Delta - \frac{3\log(K)}{2\beta}\right\}\right) + \delta_{3,\phi}(\Delta,\beta),
\end{aligned}$$

and similarly

$$\begin{aligned}
&I\!P\left(\bigcup\nolimits_{j=1}^{K}\left\{\|\boldsymbol{\phi}_j\| > z_j\right\}\right) \\
&\qquad \geq I\!P\left(\bigcup\nolimits_{j=1}^{K}\left\{\|\overline{\boldsymbol{\phi}}_j\| > z_j + \frac{3\log(K)}{2\beta} + \Delta\right\}\right) - \delta_{3,\phi}(\Delta,\beta).
\end{aligned}$$

□

The next lemma is formulated separately, since it is used for a proof of another result.

Lemma C.2 (Smooth uniform GAR)**.** *Under the conditions of Lemma C.1 it holds with dominating probability for the function $F_{\Delta,\beta}(\cdot,\boldsymbol{z})$ given in* (C.12)*:*

1.1. $I\!P\left(\bigcup_{j=1}^{K}\left\{\|\boldsymbol{\phi}_j\| > z_j\right\}\right) \leq I\!E F_{\Delta,\beta}\left(\overline{\Phi}, \boldsymbol{z} - \Delta\mathbf{1}_K\right) + \delta_{3,\phi}(\Delta,\beta)$,

1.2. $I\!P\left(\bigcup_{j=1}^{K}\left\{\|\boldsymbol{\phi}_j\| > z_j\right\}\right) \geq I\!E H_{\Delta,\beta}\left(\overline{\Phi}, \boldsymbol{z} + \frac{3\log(K)}{2\beta}\mathbf{1}_K\right) - \delta_{3,\phi}(\Delta,\beta)$;

2.1. $I\!E F_{\Delta,\beta}(\Phi,\boldsymbol{z}) \leq I\!P\left(\bigcup_{j=1}^{K}\left\{\|\boldsymbol{\phi}_j\| > z_j - \frac{3\log(K)}{2\beta}\right\}\right)$,

2.2. $I\!E F_{\Delta,\beta}(\Phi,\boldsymbol{z}) \geq I\!P\left(\bigcup_{j=1}^{K}\left\{\|\boldsymbol{\phi}_j\| > z_j + \Delta\right\}\right)$.

Proof of Lemma C.2. The first inequality 1.1 is obtained in (C.16), the second inequality 1.2 follows similarly from (C.14) and (C.15). The inequalities 2.1 and 2.2 are given in (C.13) and (C.14). □

Lemma C.3. *Let for some $\mathfrak{c}_\phi, \mathfrak{g}_1, \nu_0 > 0$ and for all $i = 1,\ldots,n$, $j = 1,\ldots,p_{\text{sum}}$*

$$\log I\!E \exp\left\{\lambda\sqrt{n}|\phi_i^j|/\mathfrak{c}_\phi\right\} \leq \lambda^2\nu_0^2/2, \quad |\lambda| < \mathfrak{g}_1,$$

here ϕ_i^j denotes the j-th coordinate of vector $\boldsymbol{\phi}_i$. Then it holds for all $i = 1,\ldots,n$ and $m, t > 0$

$$I\!P\left(\max_{1\leq j\leq p_{\text{sum}}}|\phi_i^j|^m > t\right) \leq \exp\left\{-\frac{nt^{2/m}}{2\mathfrak{c}_\phi^2\nu_0^2} + \log(p_{\text{sum}})\right\}.$$

Proof of Lemma C.3. Let us bound the $\max_j |\phi_i^j|$ using the following bound for the maximum:

$$\max_{1\le j\le p_{\mathrm{sum}}} |\phi_i^j| \le \log\Big\{\sum\nolimits_{j=1}^{p_{\mathrm{sum}}} \exp\big(|\phi_i^j|\big)\Big\}.$$

By the Lemma's condition

$$I\!E \exp\left\{\max_{1\le j\le p} \frac{\lambda\sqrt{n}}{\mathfrak{c}_\phi}|\phi_i^j|\right\} \le \exp\left(\lambda^2\nu_0^2/2 + \log p_{\mathrm{sum}}\right).$$

Thus, the statement follows from the exponential Chebyshev's inequality. □

Lemma C.4. *If for the centered random vectors* $\boldsymbol{\phi}_j \in I\!R^{p_j}$ $j = 1, \ldots, K$

$$\sup_{\substack{\boldsymbol{\gamma}\in I\!R^{p_j},\\ \|\boldsymbol{\gamma}\|\neq 0}} \log I\!E \exp\left\{\lambda \frac{\boldsymbol{\gamma}^\top \boldsymbol{\phi}_j}{\|\operatorname{Var}^{1/2}(\boldsymbol{\phi}_j)\boldsymbol{\gamma}\|}\right\} \le \nu_0^2\lambda^2/2, \qquad |\lambda| \le \mathtt{g}$$

for some constants $\nu_0 > 0$ *and* $\mathtt{g} \ge \nu_0^{-1}\max_{1\le j\le K}\sqrt{2p_j\log(K)}$, *then*

$$I\!E \max_{1\le j\le K}\left\{\|\boldsymbol{\phi}_j\|\right\} \le \mathtt{C}\nu_0 \max_{1\le j\le K}\|\operatorname{Var}^{1/2}(\boldsymbol{\phi}_j)\|\sqrt{2p_{\max}\log(K)},$$
$$\left(I\!E \max_{1\le j\le K}\left\{\|\boldsymbol{\phi}_j\|^6\right\}\right)^{1/2} \le \mathtt{C}\nu_0 \max_{1\le j\le K}\|\operatorname{Var}^{1/2}(\boldsymbol{\phi}_j)\|^3\sqrt{2p_{\max}\log(K)}(p_{\max}+6\mathtt{x}),$$

The second bound holds with probability $\ge 1 - 2\mathrm{e}^{-\mathtt{x}}$.

Proof of Lemma C.4. Let us take for each $j = 1, \ldots, K$ finite ε_j-grids $\mathrm{G}_j(\varepsilon) \subset I\!R^{p_j}$ on the $(p_j - 1)$-spheres of radius 1 s.t

$$\forall\boldsymbol{\gamma} \in I\!R^{p_j} \text{ s.t. } \|\boldsymbol{\gamma}\| = 1 \;\; \exists\boldsymbol{\gamma}_0 \in \mathrm{G}_j(\varepsilon): \;\|\boldsymbol{\gamma} - \boldsymbol{\gamma}_0\| \le \varepsilon,\; \|\boldsymbol{\gamma}_0\| = 1.$$

Then

$$\|\boldsymbol{\phi}_j\| \le (1-\varepsilon_j)^{-1}\max_{\boldsymbol{\gamma}\in\mathrm{G}_j(\varepsilon_j)}\left\{\boldsymbol{\gamma}^\top\boldsymbol{\phi}_j\right\}.$$

Hence, by inequality (C.9) and the imposed condition it holds for all

$0 < \mu < \mathtt{g}/\max_{1\le j\le K} \| \operatorname{Var}^{1/2}(\boldsymbol{\phi}_j)\|$:

$$
\begin{aligned}
\mathbb{E} \max_{1\le j\le K} \{\|\boldsymbol{\phi}_j\|\} &\le \max_{1\le j\le K} \frac{1}{1-\varepsilon_j} \mathbb{E} \max_{1\le j\le K} \max_{\boldsymbol{\gamma}\in \mathrm{G}_j(\varepsilon_j)} \left\{ \boldsymbol{\gamma}^\top \boldsymbol{\phi}_j \right\} \\
&\le \mathtt{C} \frac{1}{\mu} \mathbb{E} \log \left\{ \sum_{1\le j\le K} \sum_{\boldsymbol{\gamma}\in \mathrm{G}_j(\varepsilon_j)} \exp\left(\mu \boldsymbol{\gamma}^\top \boldsymbol{\phi}_j \right) \right\} \\
&\le \mathtt{C} \frac{1}{\mu} \log \left\{ \sum_{1\le j\le K} \sum_{\boldsymbol{\gamma}\in \mathrm{G}_j(\varepsilon_j)} \mathbb{E} \exp\left(\mu \boldsymbol{\gamma}^\top \boldsymbol{\phi}_j \right) \right\} \\
&\le \mathtt{C} \max_{1\le j\le K} \frac{\log(K \operatorname{card}\{\mathrm{G}_j(\varepsilon_j)\})}{\mu} + \mathtt{C} \frac{\mu \nu_0^2}{2} \max_{1\le j\le K} \| \operatorname{Var}(\boldsymbol{\phi}_j)\| \\
&\le \mathtt{C} \max_{1\le j\le K} \{p_j\} \frac{\log(K)}{\mu} + \mathtt{C} \frac{\mu \nu_0^2}{2} \max_{1\le j\le K} \| \operatorname{Var}(\boldsymbol{\phi}_j)\| \\
&= \mathtt{C} \nu_0 \max_{1\le j\le K} \{\sqrt{p_j}\} \max_{1\le j\le K} \| \operatorname{Var}^{1/2}(\boldsymbol{\phi}_j)\| \sqrt{2\log(K)} \\
&\text{for } \ \mu = \mathtt{C} \nu_0^{-1} \max_{1\le j\le K} \{\sqrt{p_j}\} \sqrt{2\log(K)} / \max_{1\le j\le K} \| \operatorname{Var}^{1/2}(\boldsymbol{\phi}_j)\|.
\end{aligned}
$$

For the second part of the statement we combine the first part with the result of Theorem A.3 on deviation of a random quadratic form: it holds with dominating probability for $V_{\phi_j}^2 \stackrel{\mathrm{def}}{=} \operatorname{Var} \boldsymbol{\phi}_j$

$$
\begin{aligned}
\|\boldsymbol{\phi}_j\|^2 &\le \mathfrak{Z}_{\mathtt{qf}}^2(\mathtt{x}, V_{\phi_j}) \\
&\le \operatorname{tr}(V_{\phi_j}^2) + 6\mathtt{x} \|V_{\phi_j}^2\| \le \|V_{\phi_j}^2\| (p_j + 6\mathtt{x}).
\end{aligned}
$$

□

Lemma C.5. *Let $\Gamma \in \mathbb{R}^{p_{\mathrm{sum}}}$, $\boldsymbol{\gamma}_j \in \mathbb{R}^{p_j}$ for $j = 1, \ldots, K$ are s.t. $\Gamma = \left(\boldsymbol{\gamma}_1^\top, \ldots, \boldsymbol{\gamma}_K^\top\right)^\top$, and $\boldsymbol{z} \stackrel{\mathrm{def}}{=} (z_1, \ldots, z_K)^\top$ s.t. $z_j \ge \sqrt{p_j}$, then it holds for the function $F_{\Delta,\beta}(\cdot, \boldsymbol{z})$ defined in* (C.12)*:*

$$
\begin{aligned}
&\left\| \nabla_\Gamma^2 F_{\Delta,\beta}(\Gamma, \boldsymbol{z}) \right\|_1 \le \mathtt{C}_2(\Delta, \beta) \max_{1\le j\le K} \left\{ \|\boldsymbol{\gamma}_j\|^2 \right\}, \qquad \mathtt{C}_2(\Delta, \beta) \stackrel{\mathrm{def}}{=} \mathtt{C} \left(\frac{1}{\Delta^2} + \frac{\beta}{\Delta} \right), \\
&\left\| \nabla_\Gamma^3 F_{\Delta,\beta}(\Gamma, \boldsymbol{z}) \right\|_1 \le \mathtt{C}_3(\Delta, \beta) \max_{1\le j\le K} \left\{ \|\boldsymbol{\gamma}_j\|^3 \right\}, \qquad \mathtt{C}_3(\Delta, \beta) \stackrel{\mathrm{def}}{=} \mathtt{C} \left(\frac{1}{\Delta^3} + \frac{\beta}{\Delta^2} + \frac{\beta^2}{\Delta} \right).
\end{aligned}
$$

Proof of Lemma C.5. Denote

$$
s(\Gamma) \stackrel{\mathrm{def}}{=} \sum\nolimits_{j=1}^K \exp\left(\beta \frac{\|\boldsymbol{\gamma}_j\|^2 - z_j^2}{2 z_j} \right), \quad h_\beta(s(\Gamma)) \stackrel{\mathrm{def}}{=} \beta^{-1} \log\{s(\Gamma)\}, \tag{C.17}
$$

then $F_{\beta,\Delta}(\Gamma, \boldsymbol{z}) = g\left(\Delta^{-1} h_\beta\left(s(\Gamma) \right) \right)$. Let γ^q denote the q-th coordinate of the vector

$\Gamma \in I\!R^{p_{\mathrm{sum}}}$. It holds for $q, l, b, r = 1, \ldots, p_{\mathrm{sum}}$:

$$\begin{aligned}
\frac{d}{d\gamma^q} F_{\beta,\Delta}(\Gamma, \boldsymbol{z}) &= \frac{1}{\Delta} g' \left\{ \Delta^{-1} h_\beta(s(\Gamma)) \right\} \frac{d}{d\gamma^q} h_\beta(s(\Gamma)), \\
\frac{d^2}{d\gamma^q d\gamma^l} F_{\beta,\Delta}(\Gamma, \boldsymbol{z}) &= \frac{1}{\Delta^2} g'' \left\{ \Delta^{-1} h_\beta(s(\Gamma)) \right\} \frac{d}{d\gamma^q} h_\beta(s(\Gamma)) \frac{d}{d\gamma^l} h_\beta(s(\Gamma)) \\
&\quad + \frac{1}{\Delta} g' \left\{ \Delta^{-1} h_\beta(s(\Gamma)) \right\} \frac{d^2}{d\gamma^q d\gamma^l} h_\beta(s(\Gamma)), \\
\frac{d^3}{d\gamma^q d\gamma^l d\gamma^b} F_{\beta,\Delta}(\Gamma, \boldsymbol{z}) &= \frac{1}{\Delta^3} g''' \left\{ \Delta^{-1} h_\beta(s(\Gamma)) \right\} \frac{d}{d\gamma^q} h_\beta(s(\Gamma)) \frac{d}{d\gamma^l} h_\beta(s(\Gamma)) \frac{d}{d\gamma^b} h_\beta(s(\Gamma)) \\
&\quad + \frac{1}{\Delta^2} g'' \left\{ \Delta^{-1} h_\beta(s(\Gamma)) \right\} \left\{ \frac{d^2}{d\gamma^q d\gamma^b} h_\beta(s(\Gamma)) \frac{d}{d\gamma^l} h_\beta(s(\Gamma)) \right. \\
&\quad \left. + \frac{d}{d\gamma^q} h_\beta(s(\Gamma)) \frac{d^2}{d\gamma^l d\gamma^b} h_\beta(s(\Gamma)) + \frac{d}{d\gamma^b} h_\beta(s(\Gamma)) \frac{d^2}{d\gamma^q d\gamma^l} h_\beta(s(\Gamma)) \right\} \\
&\quad + \frac{1}{\Delta} g' \left\{ \Delta^{-1} h_\beta(s(\Gamma)) \right\} \frac{d^3}{d\gamma^q d\gamma^l d\gamma^b} h_\beta(s(\Gamma)).
\end{aligned}$$

Let for $1 \le q \le p_{\mathrm{sum}}$ $j(q)$ denote an index from 1 to K s.t. the coordinate γ^q of the vector $\Gamma = \left(\boldsymbol{\gamma}_1^\top, \ldots, \boldsymbol{\gamma}_K^\top\right)^\top$ belongs to its sub-vector $\boldsymbol{\gamma}_{j(q)}$.

$$\frac{d}{d\gamma^q} h_\beta(s(\Gamma)) = \frac{1}{\beta} \frac{1}{s(\Gamma)} \frac{d}{d\gamma^q} s(\Gamma) = \frac{1}{s(\Gamma)} \frac{\gamma^q}{z_{j(q)}} \exp\left(\beta \frac{\|\boldsymbol{\gamma}_{j(q)}\|^2 - z_{j(q)}^2}{2 z_{j(q)}} \right),$$

$$\begin{aligned}
\frac{d^2}{d\gamma^q d\gamma^l} h_\beta(s(\Gamma)) &= \frac{1}{\beta} \frac{1}{s(\Gamma)} \frac{d^2}{d\gamma^q d\gamma^l} s(\Gamma) - \frac{1}{\beta} \frac{1}{s^2(\Gamma)} \frac{d}{d\gamma^q} s(\Gamma) \frac{d}{d\gamma^l} s(\Gamma) \\
&= \begin{cases}
\left\{ \frac{1}{z_{j(q)}} + \beta \left(\frac{\gamma^q}{z_{j(q)}} \right)^2 \right\} \frac{1}{s(\Gamma)} \exp\left(\beta \frac{\|\boldsymbol{\gamma}_{j(q)}\|^2 - z_{j(q)}^2}{2 z_{j(q)}} \right) & \\
\quad - \frac{\beta}{s^2(\Gamma)} \left\{ \frac{\gamma^q}{z_{j(q)}} \right\}^2 \exp\left(2\beta \frac{\|\boldsymbol{\gamma}_{j(q)}\|^2 - z_{j(q)}^2}{2 z_{j(q)}} \right), & q = l; \\
\frac{\beta}{s(\Gamma)} \frac{\gamma^q \gamma^l}{z_{j(q)}^2} \exp\left(\beta \frac{\|\boldsymbol{\gamma}_{j(q)}\|^2 - z_{j(q)}^2}{2 z_{j(q)}} \right) & \\
\quad - \frac{\beta}{s^2(\Gamma)} \frac{\gamma^q \gamma^l}{z_{j(q)}^2} \exp\left(2\beta \frac{\|\boldsymbol{\gamma}_{j(q)}\|^2 - z_{j(q)}^2}{2 z_{j(q)}} \right), & j(q) = j(l), q \ne l; \\
- \frac{\beta}{s^2(\Gamma)} \frac{\gamma^q \gamma^l}{z_{j(q)} z_{j(l)}} \exp\left(\beta \frac{\|\boldsymbol{\gamma}_{j(q)}\|^2 - z_{j(q)}^2}{2 z_{j(q)}} + \beta \frac{\|\boldsymbol{\gamma}_{j(l)}\|^2 - z_{j(l)}^2}{2 z_{j(l)}} \right), & j(q) \ne j(l).
\end{cases}
\end{aligned}$$

By definition (C.17) of $s(\Gamma)$ it holds for all $\Gamma \in I\!R^{p_{\mathrm{sum}}}$:

$$\frac{1}{s(\Gamma)} \exp\left(\beta \frac{\|\boldsymbol{\gamma}_j\|^2 - z_j^2}{2 z_j} \right) \le 1, \quad \sum_{j=1}^K \frac{1}{s(\Gamma)} \exp\left(\beta \frac{\|\boldsymbol{\gamma}_j\|^2 - z_j^2}{2 z_j} \right) = 1.$$

Therefore,

$$\begin{aligned}\sum_{q,l=1}^{p_{\mathrm{sum}}} \left| \frac{d}{d\gamma^q} h_\beta(s(\Gamma)) \frac{d}{d\gamma^l} h_\beta(s(\Gamma)) \right| &\le \left\{ \sum_{j=1}^{K} \frac{1}{s(\Gamma) z_j} \exp\left(\beta \frac{\|\boldsymbol{\gamma}_j\|^2 - z_j^2}{2 z_j} \right) \sum_{q=1}^{p_j} \gamma^q \right\}^2 \\ &\le \left| \max_{1 \le j \le K} \|\boldsymbol{\gamma}_j\| \frac{\sqrt{p_j}}{z_j} \right|^2 \\ &\le \max_{1 \le j \le K} \|\boldsymbol{\gamma}_j\|^2 \quad \text{for } z_j \ge \sqrt{p_j}.\end{aligned}$$

Similarly

$$\sum_{q,l=1}^{p_{\mathrm{sum}}} \left| \frac{d^2}{d\gamma^q d\gamma^l} h_\beta(s(\Gamma)) \right| \le \mathtt{C} \beta \max_{1 \le j \le K} \|\boldsymbol{\gamma}_j\|^2,$$

$$\sum_{q,l,b=1}^{p_{\mathrm{sum}}} \left| \frac{d^2}{d\gamma^q d\gamma^l} h_\beta(s(\Gamma)) \frac{d}{d\gamma^b} h_\beta(s(\Gamma)) + \frac{d^3}{d\gamma^q d\gamma^l d\gamma^b} h_\beta(s(\Gamma)) \right| \le \mathtt{C} \left(\beta + \beta^2 \right) \max_{1 \le j \le K} \|\boldsymbol{\gamma}_j\|^3.$$

□

C.2 Gaussian comparison

The following lemma shows how to compare the expected values of a twice differentiable function evaluated at the independent centered Gaussian vectors. This statement is used for the Gaussian comparison step in the scheme (3.6). The proof of the result is based on the Gaussian interpolation method introduced by Stein (1981) and Slepian (1962) (see also Röllin (2013) and Chernozhukov et al. (2013b) and references therein). The proof is given here in order to keep the text self-contained.

Lemma C.6 (Gaussian comparison using Slepian interpolation)**.** *Let the $\mathbb{R}^{p_{\mathrm{sum}}}$-dimensional random centered vectors $\overline{\Phi}$ and $\overline{\Psi}$ be independent and normally distributed, $f(Z) : \mathbb{R}^{p_{\mathrm{sum}}} \mapsto \mathbb{R}$ is any twice differentiable function s.t. the expected values in the expression below are bounded. Then it holds*

$$\left| \mathbb{E} f(\overline{\Phi}) - \mathbb{E} f(\overline{\Psi}) \right| \le \frac{1}{2} \left\| \operatorname{Var} \overline{\Phi} - \operatorname{Var} \overline{\Psi} \right\|_{\max} \sup_{t \in [0,1]} \left\| \mathbb{E} \nabla^2 f \left(\overline{\Phi} \sqrt{t} + \overline{\Psi} \sqrt{1-t} \right) \right\|_1.$$

Proof of Lemma C.6. Introduce for $t \in [0,1]$ the Gaussian vector process Z_t and the deterministic scalar-valued function $\varkappa(t)$:

$$\begin{aligned} Z_t &\stackrel{\mathrm{def}}{=} \overline{\Phi} \sqrt{t} + \overline{\Psi} \sqrt{1-t} \in \mathbb{R}^{p_{\mathrm{sum}}}, \\ \varkappa(t) &\stackrel{\mathrm{def}}{=} \mathbb{E} f(Z(t)), \end{aligned}$$

then $I\!E f(\overline{\Phi}) = \varkappa(1)$, $I\!E f(\overline{\Psi}) = \varkappa(0)$ and

$$\left|I\!E f(\overline{\Phi}) - I\!E f(\overline{\Psi})\right| = |\varkappa(1) - \varkappa(0)| \le \int_0^1 \left|\varkappa'(t)\right| dt.$$

Let us consider $\varkappa'(t)$:

$$\begin{aligned}\varkappa'(t) &= \frac{d}{dt} I\!E f(Z_t) = I\!E\left[\{\nabla f(Z_t)\}^\top \frac{d}{dt} Z_t\right] \\ &= \frac{1}{2\sqrt{t}} I\!E\left\{\overline{\Phi}^\top \nabla f(Z_t)\right\} - \frac{1}{2\sqrt{1-t}} I\!E\left\{\overline{\Psi}^\top \nabla f(Z_t)\right\}. \end{aligned} \tag{C.18}$$

Further we use the Gaussian integration by parts formula (see e.g Section A.6 in Talagrand (2003)): if $(x_1, \ldots, x_{p_{\text{sum}}})^\top$ is a centered Gaussian vector and $f(x_1, \ldots, x_{p_{\text{sum}}})$ is s.t. the integrals below exist, then it holds for all $j = 1, \ldots, p_{\text{sum}}$:

$$I\!E\left\{x_j f(x_1, \ldots, x_{p_{\text{sum}}})\right\} = \sum_{k=1}^{p_{\text{sum}}} I\!E(x_j x_k) I\!E\left\{\frac{d}{dx_k} f(x_1, \ldots, x_{p_{\text{sum}}})\right\}. \tag{C.19}$$

Let $\overline{\Phi}^j, \overline{\Psi}^j$ denote the j-th coordinates of $\overline{\Phi}$ and $\overline{\Psi}$. Let also $\frac{d}{d_j} f(Z_t)$ denote the partial derivative of the vectors $f(Z_t)$ w.r.t. the j-th coordinate of Z_t. Then it holds due to (C.19):

$$\begin{aligned} I\!E\left\{\overline{\Phi}^\top \nabla f(Z_t)\right\} &= \sum_{j=1}^{p_{\text{sum}}} I\!E\left\{\overline{\Phi}^j \frac{d}{d_j} f(Z_t)\right\} = \sum_{j,q=1}^{p_{\text{sum}}} I\!E\left(\overline{\Phi}^j \overline{\Phi}^q\right) I\!E\left\{\frac{d}{d\overline{\Phi}^q} \frac{d}{d_j} f(Z_t)\right\} \\ &= \sqrt{t} \sum_{j,q=1}^{p_{\text{sum}}} I\!E\left(\overline{\Phi}^j \overline{\Phi}^q\right) I\!E\left\{\frac{d^2}{d_q d_j} f(Z_t)\right\}. \end{aligned}$$

Similarly for the second term in (C.18):

$$I\!E\left\{\overline{\Psi}^\top \nabla f(Z_t)\right\} = \sqrt{1-t} \sum_{j,q=1}^{p_{\text{sum}}} I\!E\left(\overline{\Psi}^j \overline{\Psi}^q\right) I\!E\left\{\frac{d^2}{d_q d_j} f(Z_t)\right\},$$

therefore

$$\begin{aligned}\varkappa'(t) &= \frac{1}{2} \sum_{j=1}^{p_{\text{sum}}} \sum_{q=1}^{p_{\text{sum}}} \left\{I\!E\left(\overline{\Phi}^j \overline{\Phi}^q\right) - I\!E\left(\overline{\Psi}^j \overline{\Psi}^q\right)\right\} I\!E\left\{\frac{d^2}{d_q d_j} f(Z_t)\right\} \\ &\le \frac{1}{2} \left\|\operatorname{Var}\overline{\Phi} - \operatorname{Var}\overline{\Psi}\right\|_{\max} \sup_{t\in[0,1]} \left\|I\!E \nabla^2 f(Z_t)\right\|_1. \end{aligned}$$

□

C.3 Simultaneous anti-concentration for ℓ_2-norms of Gaussian vectors

Lemma C.7 (Simultaneous Gaussian anti-concentration)**.** *Let* $\left(\overline{\boldsymbol{\phi}}_1^\top, \ldots, \overline{\boldsymbol{\phi}}_K^\top\right)^\top \in I\!R^{p_{\mathrm{sum}}}$ *be centered normally distributed random vector, and* $\overline{\boldsymbol{\phi}}_j \in I\!R^{p_j}$, $j = 1, \ldots, K$. *It holds for all* $z_j \geq \sqrt{p_j}$ *and* $0 < \Delta_j \leq z_j$, $j = 1, \ldots, K$ *:*

$$I\!P\left(\bigcup\nolimits_{j=1}^K \left\{\|\overline{\boldsymbol{\phi}}_j\| > z_j\right\}\right) - I\!P\left(\bigcup\nolimits_{j=1}^K \left\{\|\overline{\boldsymbol{\phi}}_j\| > z_j + \Delta_j\right\}\right) \leq \Delta_{\mathrm{ac}}\left(\{\Delta_j\}\right),$$

where

$$\Delta_{\mathrm{ac}}\left(\{\Delta_j\}\right) \leq \mathtt{C}\left\{\varkappa\sqrt{1 \vee \log(K/2)} + \mathtt{C}\max_{1\leq j\leq K}\{\Delta_j\}\sqrt{\max_{1\leq j\leq K}\log(2z_j/\Delta_j)}\right\},$$

and $\varkappa \stackrel{\mathrm{def}}{=} \max_{1\leq j\leq K}\{\Delta_j/z_j\} \leq 1$ *is a deterministic positive constant. An explicit definition of* $\Delta_{\mathrm{ac}}\left(\{\Delta_j\}\right)$ *is given in* (C.22)*.*

Proof of Lemma C.7.

$$\begin{aligned}
&I\!P\left(\bigcup\nolimits_{j=1}^K \left\{\|\overline{\boldsymbol{\phi}}_j\| > z_j\right\}\right) - I\!P\left(\bigcup\nolimits_{j=1}^K \left\{\|\overline{\boldsymbol{\phi}}_j\| > z_j + \Delta_j\right\}\right) \\
&\quad \leq I\!P\left(\bigcup\nolimits_{j=1}^K \left\{\|\overline{\boldsymbol{\phi}}_j\| z_j^{-1} - 1 > 0\right\}\right) - I\!P\left(\bigcup\nolimits_{j=1}^K \left\{\|\overline{\boldsymbol{\phi}}_j\| z_j^{-1} - 1 > \varkappa\right\}\right) \\
&\quad = I\!P\left(\max_{1\leq j\leq K}\left\{\|\overline{\boldsymbol{\phi}}_j\| z_j^{-1} - 1\right\} > 0\right) - I\!P\left(\max_{1\leq j\leq K}\left\{\|\overline{\boldsymbol{\phi}}_j\| z_j^{-1} - 1\right\} > \varkappa\right) \\
&\quad \leq I\!P\left(0 \leq \max_{1\leq j\leq K}\left\{\|\overline{\boldsymbol{\phi}}_j\| z_j^{-1} - 1\right\} \leq \varkappa\right). \qquad \text{(C.20)}
\end{aligned}$$

It holds

$$\|\overline{\boldsymbol{\phi}}_j\| = \sup_{\substack{\boldsymbol{\gamma}\in I\!R^{p_j},\\ \|\boldsymbol{\gamma}\|=1}} \left\{\boldsymbol{\gamma}^\top \overline{\boldsymbol{\phi}}_j\right\}.$$

Let $\mathrm{G}_j(\varepsilon_j) \subset I\!R^{p_j}$ (for $1 \leq j \leq K$) denote a finite ε_j-net on $(p_j - 1)$-sphere of radius 1 :

$$\forall \boldsymbol{\gamma} \in I\!R^{p_j} \text{ s.t. } \|\boldsymbol{\gamma}\| = 1 \;\; \exists \boldsymbol{\gamma}_0 \in \mathrm{G}_j(\varepsilon_j) : \; \|\boldsymbol{\gamma} - \boldsymbol{\gamma}_0\| \leq \varepsilon_j, \; \|\boldsymbol{\gamma}_0\| = 1.$$

This implies for all $j = 1, \ldots, K$

$$(1 - \varepsilon_j)\|\overline{\boldsymbol{\phi}}_j\| \leq \max_{\boldsymbol{\gamma}\in \mathrm{G}_j(\varepsilon_j)} \left\{\boldsymbol{\gamma}^\top \overline{\boldsymbol{\phi}}_j\right\} \leq \|\overline{\boldsymbol{\phi}}_j\|.$$

Let us take $\varepsilon_1, \ldots, \varepsilon_K > 0$ s.t. $\forall\, j = 1, \ldots, K$

$$\varepsilon_j \|\overline{\boldsymbol{\phi}}_j\| z_j^{-1} \leq \varkappa, \qquad \text{(C.21)}$$

then

$$0 \le \max_{1\le j\le K}\left\{\frac{\|\overline{\boldsymbol{\phi}}_j\|}{z_j}\right\} - \max_{1\le j\le K}\max_{\boldsymbol{\gamma}\in \mathrm{G}_j(\varepsilon_j)}\left\{\frac{\boldsymbol{\gamma}^\top\overline{\boldsymbol{\phi}}_j}{z_j}\right\} \le \varkappa,$$

and the inequality (C.20) continues as

$$\begin{aligned}&I\!P\left(0 \le \max_{1\le j\le K}\left\{\|\overline{\boldsymbol{\phi}}_j\| z_j^{-1} - 1\right\} \le \varkappa\right)\\ &\qquad \le I\!P\left(\left|\max_{1\le j\le K}\sup_{\boldsymbol{\gamma}\in \mathrm{G}_j(\varepsilon_j)}\left\{\frac{\boldsymbol{\gamma}^\top\overline{\boldsymbol{\phi}}_j}{z_j}\right\} - 1\right| \le \varkappa\right).\end{aligned}$$

The random values $\boldsymbol{\gamma}^\top\overline{\boldsymbol{\phi}}_j z_j^{-1} \sim \mathcal{N}(0, z_j^{-2}\operatorname{Var}\{\boldsymbol{\gamma}^\top\overline{\boldsymbol{\phi}}_j\})$. The anti-concentration inequality by Chernozhukov et al. (2014c) for the maximum of a centered high-dimensional Gaussian vector (see Theorem C.1 below), applied to $\max_{1\le j\le K}\sup_{\boldsymbol{\gamma}\in \mathrm{G}_j(\varepsilon_j)}\left\{\boldsymbol{\gamma}^\top\overline{\boldsymbol{\phi}}_j z_j^{-1}\right\}$, implies

$$\begin{aligned}&I\!P\left(\left|\max_{1\le j\le K}\sup_{\boldsymbol{\gamma}\in \mathrm{G}_j(\varepsilon_j)}\left\{\frac{\boldsymbol{\gamma}^\top\overline{\boldsymbol{\phi}}_j}{z_j}\right\} - 1\right| \le \varkappa\right)\\ &\qquad \le \Delta_{\mathrm{ac}} \stackrel{\mathrm{def}}{=} \mathtt{C}_{\mathrm{ac}}\varkappa\sqrt{1 \vee \log\left(\varkappa^{-1}\sum\nolimits_{j=1}^{K}\{2/\varepsilon_j\}^{p_j}\right)}, \qquad \text{(C.22)}\end{aligned}$$

where the constant $\mathtt{C}_{\mathrm{ac}}$ depends on min and max of $\operatorname{Var}\{\boldsymbol{\gamma}^\top\overline{\boldsymbol{\phi}}_j z_j^{-1}\} \le I\!E\|\overline{\boldsymbol{\phi}}_j\|^2 z_j^{-2} \le 1$; the sum $\sum_{j=1}^{K}\{2/\varepsilon_j\}^{p_j}$ is proportional to cardinality of the set $\{\boldsymbol{\gamma}^\top\overline{\boldsymbol{\phi}}_j z_j^{-1}, \boldsymbol{\gamma} \in \mathrm{G}_j(\varepsilon_j),\, j = 1,\ldots,K\}$. If one takes $\varepsilon_j = 2\mathtt{C}\{\Delta_j/(2z_j)\}^{\frac{p_{\min}+1}{p_j+1}}$, then (C.21) holds with exponentially high probability due to Gaussianity of the vectors $\overline{\boldsymbol{\phi}}_j$ and Theorem 1.2 in Spokoiny (2012b), hence

$$\begin{aligned}\Delta_{\mathrm{ac}} &\le \mathtt{C}_{\mathrm{ac}}\varkappa\sqrt{1 \vee \mathtt{C}\log\left(\frac{1}{2}\sum\nolimits_{j=1}^{K}\{2/\varepsilon_j\}^{p_j+1}\right)}\\ &\le \mathtt{C}_{\mathrm{ac}}\left\{\varkappa\sqrt{1 \vee \log(K/2)} + \mathtt{C}\max_{1\le j\le K}\{\Delta_j\}\sqrt{\max_{1\le j\le K}\log(2z_j/\Delta_j)}\right\}. \qquad \text{(C.23)}\end{aligned}$$

□

Theorem C.1 (Anti-concentration inequality for maxima of a Gaussian random vector, Chernozhukov et al. (2014c)). *Let* $(X_1,\ldots,X_p)^\top$ *be a centered Gaussian random vector with* $\sigma_j^2 \stackrel{\mathrm{def}}{=} I\!EX_j^2 > 0$ *for all* $1 \le j \le p$. *Let* $\underline{\sigma} \stackrel{\mathrm{def}}{=} \min_{1\le j\le p}\sigma_j$, $\overline{\sigma} \stackrel{\mathrm{def}}{=} \max_{1\le j\le p}\sigma_j$. *Then for every* $\epsilon > 0$

$$\sup_{x\in I\!R} I\!P\left(\left|\max_{1\le j\le p} X_j - x\right| \le \epsilon\right) \le \mathtt{C}_{\mathrm{ac}}\epsilon\sqrt{1 \vee \log(p/\epsilon)},$$

where $\mathtt{C}_{\mathrm{ac}}$ *depends only on* $\underline{\sigma}$ *and* $\overline{\sigma}$. *When the variances are all equal, namely* $\underline{\sigma} = \overline{\sigma} = \sigma$, $\log(p/\epsilon)$ *on the right side can be replaced by* $\log p$.

C.4 Proof of Proposition C.1

Proof of Proposition C.1. Let $\Phi \stackrel{\text{def}}{=} \left(\boldsymbol{\phi}_1^\top, \ldots, \boldsymbol{\phi}_K^\top\right)^\top \in I\!R^{p_{\text{sum}}}$ for $p_{\text{sum}} \stackrel{\text{def}}{=} p_1 + \cdots + p_K$ (as in (C.5)), and similarly $\Psi \stackrel{\text{def}}{=} \left(\boldsymbol{\psi}_1^\top, \ldots, \boldsymbol{\psi}_K^\top\right)^\top \in I\!R^{p_{\text{sum}}}$. Let also $\overline{\Phi} \sim \mathcal{N}(0, \operatorname{Var}\Phi)$ and $\overline{\Psi} \sim \mathcal{N}(0, \operatorname{Var}\Psi)$. Introduce the following value, which comes from Lemma C.6 on Gaussian comparison:

$$\delta_2(\Delta, \beta) \stackrel{\text{def}}{=} \mathtt{C}_2(\Delta, \beta) \max_{1 \le j \le K} \sup_{t \in [0,1]} \left\{ I\!E \|\overline{\boldsymbol{\phi}}_j \sqrt{t} + \overline{\boldsymbol{\psi}}_j \sqrt{1-t}\|^2 \right\}$$
$$\le \mathtt{C}_2(\Delta, \beta) \max_{1 \le j \le K} \max\left\{ \operatorname{tr}\operatorname{Var}(\overline{\boldsymbol{\phi}}_j), \operatorname{tr}\operatorname{Var}(\overline{\boldsymbol{\psi}}_j) \right\}. \tag{C.24}$$

It holds

$$I\!P\left(\bigcup\nolimits_{j=1}^K \left\{ \|\boldsymbol{\phi}_j\| > z_j \right\} \right)$$
$$\stackrel{\text{by L. C.2}}{\ge} I\!E H_{\Delta,\beta}\left(\overline{\Phi}, \boldsymbol{z} + \frac{3\log(K)}{2\beta} \mathbf{1}_K \right) - \delta_{3,\phi}(\Delta, \beta)$$
$$\stackrel{\text{by L. C.6, C.5}}{\ge} I\!E H_{\Delta,\beta}\left(\overline{\Psi}, \boldsymbol{z} + \frac{3\log(K)}{2\beta} \mathbf{1}_K \right) - \frac{1}{2}\delta_\Sigma^2 \delta_2(\Delta, \beta) - \delta_{3,\phi}(\Delta, \beta)$$
$$\stackrel{\text{by L. C.2}}{\ge} I\!P\left(\bigcup\nolimits_{j=1}^K \left\{ \|\overline{\boldsymbol{\psi}}_j\| > z_j + \Delta + \frac{3\log(K)}{2\beta} \right\} \right) - \frac{1}{2}\delta_\Sigma^2 \delta_2(\Delta, \beta) - \delta_{3,\phi}(\Delta, \beta)$$
$$\stackrel{\text{by L. C.7}}{\ge} I\!P\left(\bigcup\nolimits_{j=1}^K \left\{ \|\overline{\boldsymbol{\psi}}_j\| > z_j - \delta_{z_j} - \Delta \right\} \right) - \frac{1}{2}\delta_\Sigma^2 \delta_2(\Delta, \beta) - \delta_{3,\phi}(\Delta, \beta)$$
$$- 2\Delta_{\text{ac}}\left(\left\{ \delta_{z_j} \right\} + 2\Delta + \frac{3\log(K)}{\beta} \right) \tag{C.25}$$
$$\stackrel{\text{by L. C.1}}{\ge} I\!P\left(\bigcup\nolimits_{j=1}^K \left\{ \|\boldsymbol{\psi}_j\| > z_j - \delta_{z_j} \right\} \right) - \frac{1}{2}\delta_\Sigma^2 \delta_2(\Delta, \beta) \tag{C.26}$$
$$- \delta_{3,\phi}(\Delta, \beta) - \delta_{3,\psi}(\Delta, \beta) - 2\Delta_{\text{ac}}\left(\left\{ \delta_{z_j} \right\} + 2\Delta + \frac{3\log(K)}{\beta} \right),$$

where $\delta_{3,\psi}(\Delta, \beta)$ is defined similarly to $\delta_{3,\phi}(\Delta, \beta)$ in (C.15):

$$\delta_{3,\psi}(\Delta, \beta) \stackrel{\text{def}}{=} \frac{\mathtt{C}_3(\Delta, \beta)}{3} \frac{p_{\max}^{3/2}}{n^{1/2}} \log^{1/2}(K) \log^{3/2}(n p_{\text{sum}}) \left(2\nu_0^2 \mathfrak{c}_\psi^2 \lambda_{\psi,\max}^2 \right)^{3/2}. \tag{C.27}$$

By Lemma C.7 inequality (C.25) requires the following: $\delta_{z_j} + 2\Delta + \frac{3\log(K)}{\beta} \le z_j$ .The bound in the inverse direction is derived similarly. Denote the approximating error term obtained in (C.26) as

$$\Delta_{\ell_2} \stackrel{\text{def}}{=} \frac{1}{2}\delta_\Sigma^2 \delta_2(\Delta, \beta) + \delta_{3,\phi}(\Delta, \beta) + \delta_{3,\psi}(\Delta, \beta) + 2\Delta_{\text{ac}}\left(\left\{ \delta_{z_j} \right\} + 2\Delta + \frac{3\log(K)}{\beta} \right).$$

Consider this term in more details, by inequality (C.23)

$$\Delta_{\text{ac}}\left(\left\{ \delta_{z_j} \right\} + 2\Delta + \frac{3\log(K)}{\beta} \right) \le \max_{1 \le j \le K} \left(\delta_{z_j} + 2\Delta + \frac{3\log(K)}{\beta} \right)$$
$$\times \left\{ \mathtt{C} \frac{\log^{1/2}(K)}{z_j} + \log^{1/2}(2 z_{\max}) - \log^{1/2}\left(\delta_{z_j} + 2\Delta + \frac{3\log(K)}{\beta} \right) \right\}.$$

Let us take $\beta = \frac{\log(K)}{\Delta}$, then

$$\begin{aligned}
\Delta_{\text{ac}} &\leq 5\mathtt{C}\Delta\frac{\log^{1/2}(K)}{z_{\min}} + \mathtt{C}\max_{1\leq j\leq K}\frac{\delta_{z_j}}{z_j}\log^{1/2}(K) \\
&+ \mathtt{C}\left(5\Delta+\delta_{z,\max}\right)\left(\log^{1/2}\left(2z_{\max}\right)+\sqrt{-\log\left(\delta_{z,\min}+5\Delta\right)}\right), \\
&\leq 5\mathtt{C}\Delta\frac{\log^{1/2}(K)}{z_{\min}} + \mathtt{C}\max_{1\leq j\leq K}\frac{\delta_{z_j}}{z_j}\log^{1/2}(K) \\
&+ 2\mathtt{C}\left(5\Delta+\delta_{z,\max}\right)\sqrt{-\log\left(\delta_{z,\min}+5\Delta\right)} \\
&\leq 5\mathtt{C}\Delta\frac{\log^{1/2}(K)}{z_{\min}} + \mathtt{C}\max_{1\leq j\leq K}\frac{\delta_{z_j}}{z_j}\log^{1/2}(K) + 2\mathtt{C}\left(5\Delta+\delta_{z,\max}\right)\sqrt{-\log\left(5\Delta\right)} \\
&\leq 5\mathtt{C}\Delta\Big\{\frac{\log^{1/2}(K)}{z_{\min}} + 2.4\log^{1/2}\big(5n^{1/2}\big)\Big\} + \mathtt{C}\max_{1\leq j\leq K}\frac{\delta_{z_j}}{z_j}\log^{1/2}(K) \\
&\leq 6\mathtt{C}\Delta\Big\{\frac{\log^{1/2}(K)}{z_{\min}} + 0.4\log^{1/2}\big(5n^{1/2}\big)\Big\}, \qquad \text{(C.28)}
\end{aligned}$$

where the second inequality holds for $\delta_{z,\min}+5\Delta \leq 1/(2z_{\max})$, and the last one holds for $\delta_{z,\max}\leq\Delta$ and $\Delta \geq n^{-1/2}$.

$$\begin{aligned}
\delta_{3,\phi}(\Delta,\beta)+\delta_{3,\psi}(\Delta,\beta) &\overset{\text{by (C.27)}}{\leq} \mathtt{C}\frac{\log^{5/2}(K)}{\Delta^3}\frac{p_{\max}^{3/2}}{n^{1/2}}\log^{3/2}(np_{\text{sum}})\left(\lambda_{\phi,\max}^3+\lambda_{\psi,\max}^3\right), \qquad \text{(C.29)} \\
\delta_{\Sigma}\delta_2(\Delta,\beta) &\overset{\text{by (C.24)}}{\leq} \mathtt{C}\delta_{\Sigma}^2\frac{\log(K)}{\Delta^2}\max_{1\leq j\leq K}\max\left\{\operatorname{tr}\operatorname{Var}(\overline{\boldsymbol{\phi}}_j),\operatorname{tr}\operatorname{Var}(\overline{\boldsymbol{\psi}}_j)\right\} \\
&\leq \mathtt{C}\delta_{\Sigma}^2\frac{\log(K)}{\Delta^2}p_{\max}\max\left\{\lambda_{\phi,\max}^2,\lambda_{\psi,\max}^2\right\}.
\end{aligned}$$

After minimizing the sum of the expressions (C.28) and (C.29) w.r.t Δ, we have

$$\begin{aligned}
\Delta_{\ell_2} &\leq 12.5\mathtt{C}\left(\frac{p_{\max}^3}{n}\right)^{1/8}\log^{9/8}(K)\log^{3/8}(np_{\text{sum}})\max\left\{\lambda_{\phi,\max},\lambda_{\psi,\max}\right\}^{3/4} \\
&+ 3.2\mathtt{C}\delta_{\Sigma}^2 p_{\max}z_{\min}^{1/2}\left(\frac{p_{\max}^3}{n}\right)^{1/4}\log^2(K)\log^{3/4}(np_{\text{sum}})\max\left\{\lambda_{\phi,\max},\lambda_{\psi,\max}\right\}^{7/2} \\
&\leq 25\mathtt{C}\left(\frac{p_{\max}^3}{n}\right)^{1/8}\log^{9/8}(K)\log^{3/8}(np_{\text{sum}})\max\left\{\lambda_{\phi,\max},\lambda_{\psi,\max}\right\}^{3/4},
\end{aligned}$$

where the last inequality holds for

$$\delta_{\Sigma}^2 \leq 4\mathtt{C}p_{\max}^{-1}z_{\min}^{-1/2}\left(\frac{p_{\max}^3}{n}\right)^{-1/8}\log^{-7/8}(K)\log^{-3/8}(np_{\text{sum}})\left(\max\left\{\lambda_{\phi,\max},\lambda_{\psi,\max}\right\}\right)^{-11/4}.$$

□

Appendix D

Proofs of the main results

D.1 Proofs for Chapter 2

D.1.1 Proofs of Theorems 2.1 – 2.3

In order to justify theoretically the multiplier bootstrap procedure it has to be shown that the approximating terms $\|\boldsymbol{\xi}\|$ and $\|\boldsymbol{\xi}^{\circ}(\widetilde{\boldsymbol{\theta}})\|$ from the Wilks Theorems A.2 and A.4 have nearly the same distributions. By Lemma A.2 the random values $\|\boldsymbol{\xi}^{\circ}(\boldsymbol{\theta}^*)\|$ and $\|\boldsymbol{\xi}^{\circ}(\widetilde{\boldsymbol{\theta}})\|$ are close to each other within the error term $\leq \mathtt{C}(p+\mathtt{x})\sqrt{\mathtt{x}/n}$ with exponentially high probability, therefore, it is sufficient to compare the distributions of $\|\boldsymbol{\xi}^{\circ}(\boldsymbol{\theta}^*)\|$ and $\|\boldsymbol{\xi}\|$. This is done in Proposition D.1 using the results on Gaussian approximation for Euclidean norms from Section B.

Let us introduce the multivariate normal vectors similarly to (B.3):

$$\overline{\boldsymbol{\xi}} \sim \mathcal{N}(0, \operatorname{Var}\boldsymbol{\xi}), \quad \overline{\boldsymbol{\xi}}^{\circ}(\boldsymbol{\theta}^*) \sim \mathcal{N}(0, \operatorname{Var}^{\circ}\{\boldsymbol{\xi}^{\circ}(\boldsymbol{\theta}^*)\}). \tag{D.1}$$

Let us also represent the vectors $\boldsymbol{\xi}$ and $\boldsymbol{\xi}^{\circ}(\boldsymbol{\theta}^*)$ as sums of the marginal score vectors $\boldsymbol{\xi}_i$ and $\boldsymbol{\xi}_i^{\circ}(\boldsymbol{\theta}^*)$ s.t. $I\!\!E\boldsymbol{\xi}_i = I\!\!E^{\circ}\boldsymbol{\xi}_i^{\circ} = 0$:

$$\begin{aligned} \boldsymbol{\xi}_i &\stackrel{\text{def}}{=} D_0^{-1}\left\{\nabla_{\boldsymbol{\theta}}\ell_i(\boldsymbol{\theta}^*) - \nabla_{\boldsymbol{\theta}} I\!\!E\ell_i(\boldsymbol{\theta}^*)\right\}, \\ \boldsymbol{\xi}_i^{\circ}(\boldsymbol{\theta}^*) &\stackrel{\text{def}}{=} D_0^{-1}\nabla_{\boldsymbol{\theta}}\ell_i(\boldsymbol{\theta}^*)\{u_i - 1\}. \end{aligned}$$

Their Gaussian analogs are

$$\overline{\boldsymbol{\xi}}_i \sim \mathcal{N}(0, \operatorname{Var}\boldsymbol{\xi}_i) \quad \text{and} \quad \overline{\boldsymbol{\xi}}_i^{\circ} \sim \mathcal{N}(0, \operatorname{Var}^{\circ}\{\boldsymbol{\xi}_i^{\circ}(\boldsymbol{\theta}^*)\}).$$

Similarly to (B.4) denote

$$\begin{aligned} \delta_n &\stackrel{\text{def}}{=} \frac{1}{2}\sum\nolimits_{i=1}^{n} I\!\!E\left(\|\boldsymbol{\xi}_i\|^3 + \|\overline{\boldsymbol{\xi}}_i\|^3\right), \\ \breve{\delta}_n &\stackrel{\text{def}}{=} \frac{1}{2}\sum\nolimits_{i=1}^{n} I\!\!E^{\circ}\left(\|\boldsymbol{\xi}_i^{\circ}(\boldsymbol{\theta}^*)\|^3 + \|\overline{\boldsymbol{\xi}}_i^{\circ}(\boldsymbol{\theta}^*)\|^3\right). \end{aligned} \tag{D.2}$$

Proposition D.1 (Closeness of the c.d.f. of $\|\boldsymbol{\xi}\|$ and $\|\boldsymbol{\xi}^{\circ}(\boldsymbol{\theta}^*)\|$). *Let conditions* **(SmB)** *and* **(SD_1)** *be fulfilled. Let also* $z, \overline{z} \geq \max\{2, \sqrt{p}\}$ *and* $|z - \overline{z}| \leq \delta_z$ *for some* $\delta_z \geq 0$. *Then it holds for all* $0 < \Delta \leq 0.22$ *with probability* $\geq 1 - \mathrm{e}^{-\mathtt{x}}$:

$$\begin{aligned}\Big| I\!P\left(\|\boldsymbol{\xi}\| > z\right) - I\!P^{\circ}\left(\|\boldsymbol{\xi}^{\circ}(\boldsymbol{\theta}^*)\| > \overline{z}\right)\Big| \\ \leq 16\Delta^{-3}\left(\delta_n + \breve{\delta}_n\right) + \frac{2\Delta + \delta_z}{\sqrt{2}} + \frac{\sqrt{p}}{2}\frac{\delta^2_{\mathcal{V}}(\mathtt{x}) + \delta^2_{\mathrm{smb}}}{1 - \delta^2_{\mathcal{V}}(\mathtt{x})} \\ \text{for } \delta^2_{\mathcal{V}}(\mathtt{x}) \leq 1/4.\end{aligned}$$

Moreover, if $\max\{\delta_n^{1/4}, \breve{\delta}_n^{1/4}\} \leq 0.077$, *then*

$$\left| I\!P\left(\|\boldsymbol{\xi}\| > z\right) - I\!P^{\circ}\left(\|\boldsymbol{\xi}^{\circ}(\boldsymbol{\theta}^*)\| > \overline{z}\right)\right| \tag{D.3}$$

$$\leq 2.71\big(\delta_n^{1/4} + \breve{\delta}_n^{1/4}\big) + \frac{\delta_z}{\sqrt{2}} + \frac{2\sqrt{p}}{3}\left(\delta^2_{\mathcal{V}}(\mathtt{x}) + \delta^2_{\mathrm{smb}}\right). \tag{D.4}$$

Proof of Proposition D.1. We use Theorem B.1 taking $\boldsymbol{\phi} := \boldsymbol{\xi}$ and $\boldsymbol{\psi} := \boldsymbol{\xi}^{\circ}(\boldsymbol{\theta}^*)$. Let us check that the conditions (B.5) on the covariance matrices are fulfilled. By definitions (1.9), (1.10) and (A.25)

$$\begin{aligned}\operatorname{Var}\boldsymbol{\xi} &= D_0^{-1}H_0^2D_0^{-1} - D_0^{-1}B_0^2D_0^{-1}, \\ \operatorname{Var}^{\circ}\left\{\boldsymbol{\xi}^{\circ}(\boldsymbol{\theta}^*)\right\} &= D_0^{-1}\mathcal{V}^2(\boldsymbol{\theta}^*)D_0^{-1}.\end{aligned}$$

Due to Theorem D.1 by Tropp (2012) (see Section D.1.4) it holds with probability $\geq 1 - \mathrm{e}^{-\mathtt{x}}$

$$\|H_0^{-1}\mathcal{V}^2(\boldsymbol{\theta}^*)H_0^{-1} - \boldsymbol{I}_p\| \leq \delta^2_{\mathcal{V}}(\mathtt{x}), \tag{D.5}$$

therefore, by Cauchy-Schwarz inequality

$$\|\mathcal{V}^{-1}(\boldsymbol{\theta}^*)H_0^2\mathcal{V}^{-1}(\boldsymbol{\theta}^*) - \boldsymbol{I}_p\| \leq \delta^2_{\mathcal{V}}(\mathtt{x})(1 - \delta^2_{\mathcal{V}}(\mathtt{x}))^{-1}.$$

Condition **(SmB)** says that $\|H_0^{-1}B_0^2H_0^{-1}\| \leq \delta^2_{\mathrm{smb}}$, therefore, by the triangle inequality it holds:

$$\begin{aligned}\left\|\left[\operatorname{Var}^{\circ}\left\{\boldsymbol{\xi}^{\circ}(\boldsymbol{\theta}^*)\right\}\right]^{-1/2}\operatorname{Var}\{\boldsymbol{\xi}\}\left[\operatorname{Var}^{\circ}\left\{\boldsymbol{\xi}^{\circ}(\boldsymbol{\theta}^*)\right\}\right]^{-1/2} - \boldsymbol{I}_p\right\| &\leq \frac{\delta^2_{\mathcal{V}}(\mathtt{x}) + \delta^2_{\mathrm{smb}}}{1 - \delta^2_{\mathcal{V}}(\mathtt{x})} \\ &\leq 1/2 \\ \text{for } \delta^2_{\mathrm{smb}} \leq 1/8,\ \delta^2_{\mathcal{V}}(\mathtt{x}) &\leq 1/4.\end{aligned}$$

□

The following lemma is used further for the proof of Theorem 2.1.

Lemma D.1 (Anti-concentration inequality for the likelihood ratio)**.** *It holds with probability* $\geq 1 - 5\mathrm{e}^{-\mathtt{x}}$ *for* $z \geq \max\{2, \sqrt{p}\}$, $\delta_z \geq 0$ *and* $\delta_n^{1/4} \leq 0.077$

$$\mathbb{P}\Big(\sqrt{2\big\{L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*)\big\}} > \mathfrak{z}\Big) \leq \mathbb{P}\Big(\sqrt{2\big\{L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*)\big\}} > \mathfrak{z} + \delta_z\Big) + \Delta_{\mathrm{LR}} + \delta_z/\sqrt{2}.$$

for $\Delta_{\mathrm{LR}} \stackrel{\mathrm{def}}{=} 5.42\delta_n^{1/4} + \sqrt{2}\Delta_{\mathrm{W}}(\mathtt{r}_0, \mathtt{x})$

Proof of Lemma D.1. It holds on a random set of probability $\geq 1 - 12\mathrm{e}^{-\mathtt{x}}$:

$$\begin{aligned}
&\mathbb{P}\Big(\sqrt{2\big\{L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*)\big\}} > \mathfrak{z}\Big) \\
&\stackrel{(\text{Th. } A.2)}{\leq} \mathbb{P}\left(\|\boldsymbol{\xi}\| > \mathfrak{z} - \Delta_{\mathrm{W}}(\mathtt{r}_0, \mathtt{x})\right) \\
&\stackrel{(\text{Th. } B.1)}{\leq} \mathbb{P}\left(\|\overline{\boldsymbol{\xi}}\| > \mathfrak{z} - \Delta_{\mathrm{W}}(\mathtt{r}_0, \mathtt{x})\right) + 2.71\delta_n^{1/4} \\
&\stackrel{(\text{Th. } B.1)}{\leq} \mathbb{P}\left(\|\boldsymbol{\xi}\| > \mathfrak{z} + \Delta_{\mathrm{W}}(\mathtt{r}_0, \mathtt{x}) + \delta_z\right) + 5.42\delta_n^{1/4} + \frac{1}{\sqrt{2}}\left(2\Delta_{\mathrm{W}}(\mathtt{r}_0, \mathtt{x}) + \delta_z\right) \\
&\stackrel{(\text{Th. } A.2)}{\leq} \mathbb{P}\Big(\sqrt{2\big\{L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*)\big\}} > \mathfrak{z} + \delta_z\Big) + 5.42\delta_n^{1/4} + \frac{1}{\sqrt{2}}\left(2\Delta_{\mathrm{W}}(\mathtt{r}_0, \mathtt{x}) + \delta_z\right).
\end{aligned}$$

□

Now we are ready to collect all the obtained bounds together for the following

Proofs of Theorems 2.1 and 2.2. On a random set of probability $\geq 1 - 12\mathrm{e}^{-\mathtt{x}}$ it holds:

$$\begin{aligned}
&\mathbb{P}^\circ\left(\sqrt{2\Big\{L^\circ(\widetilde{\boldsymbol{\theta}}^\circ) - L^\circ(\widetilde{\boldsymbol{\theta}})\Big\}} > \mathfrak{z}\right) \\
&\stackrel{(\text{Th. } A.4)}{\geq} \mathbb{P}^\circ\left(\|\boldsymbol{\xi}^\circ(\widetilde{\boldsymbol{\theta}})\| > \mathfrak{z} + \Delta^\circ_{\mathrm{W}}(\mathtt{r}_0, \mathtt{x})\right) \\
&\stackrel{(\text{L. } A.2)}{\geq} \mathbb{P}^\circ\left(\|\boldsymbol{\xi}^\circ(\boldsymbol{\theta}^*)\| > \mathfrak{z} + \Delta^\circ_{\mathrm{W}}(\mathtt{r}_0, \mathtt{x}) + \Delta^\circ_{\xi}(\mathtt{r}_0, \mathtt{x})\right) && \text{(D.6)} \\
&\stackrel{(\text{Prop. } D.1)}{\geq} \mathbb{P}\left(\|\boldsymbol{\xi}\| > \mathfrak{z} - \Delta_{\mathrm{W}}(\mathtt{r}_0, \mathtt{x})\right) - \Delta_{\mathrm{full}} && \text{(D.7)} \\
&\stackrel{(\text{Th. } A.2)}{\geq} \mathbb{P}\left(\sqrt{2\big\{L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*)\big\}} > \mathfrak{z}\right) - \Delta_{\mathrm{full}}, && \text{(D.8)}
\end{aligned}$$

where the value Δ_{full} comes from the bound (D.4) with $\delta_z := \Delta_{\mathrm{W}}(\mathtt{r}_0, \mathtt{x}) + \Delta^\circ_{\mathrm{W}}(\mathtt{r}_0, \mathtt{x}) + \Delta^\circ_{\xi}(\mathtt{r}_0, \mathtt{x})$:

$$\begin{aligned}
\Delta_{\mathrm{full}} \stackrel{\mathrm{def}}{=}\ & 2.71\big(\delta_n^{1/4} + \breve{\delta}_n^{1/4}\big) + \frac{2\sqrt{p}}{3}\left(\delta^2_{\mathcal{V}}(\mathtt{x}) + \delta^2_{\mathrm{smb}}\right) \\
&+ \left\{\Delta_{\mathrm{W}}(\mathtt{r}_0, \mathtt{x}) + \Delta^\circ_{\mathrm{W}}(\mathtt{r}_0, \mathtt{x}) + \Delta^\circ_{\xi}(\mathtt{r}_0, \mathtt{x})\right\}/\sqrt{2}
\end{aligned} \tag{D.9}$$

By the similar arguments in the inverse direction we obtain the following inequality:

$$\left|\mathbb{P}\Big(\sqrt{2\big\{L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*)\big\}} > \mathfrak{z}\Big) - \mathbb{P}^\circ\Big(\sqrt{2\big\{L^\circ(\widetilde{\boldsymbol{\theta}}^\circ) - L^\circ(\widetilde{\boldsymbol{\theta}})\big\}} > \mathfrak{z}\Big)\right| \leq \Delta_{\mathrm{full}}. \tag{D.10}$$

Notice that inequality (D.4) from Proposition D.1, that we use here, requires $\max\{\delta_n^{1/4}, \breve{\delta}_n^{1/4}\} \leq 0.077$.

Let us quantify, how the error term Δ_{full} depends on p and n. In the case A.3.1 random vectors $\boldsymbol{\xi}_i$ and $\boldsymbol{\xi}_i^\circ(\boldsymbol{\theta}^*)$ satisfy the conditions of Theorems A.3 and A.6 correspondingly. Hence $\|\boldsymbol{\xi}_i\|, \|\boldsymbol{\xi}_i^\circ(\boldsymbol{\theta}^*)\| \leq \mathtt{C}\sqrt{(p+\mathtt{x})/n}$ and $\delta_n, \breve{\delta}_n \leq \mathtt{C}\sqrt{(p+\mathtt{x})^3/n}$. Finally we have in the case A.3.1

$$\Delta_{\text{full}} \leq \mathtt{C}\left\{\frac{(p+\mathtt{x})^3}{n}\right\}^{1/8} + \mathtt{C}\,\frac{p+\mathtt{x}}{\sqrt{n}}\sqrt{\mathtt{x}} + \mathtt{C}\,\frac{p+\mathtt{x}}{\sqrt{n}}. \tag{D.11}$$

Now let us proof Theorem 2.2. It holds with probability $\geq 1 - 12\mathrm{e}^{-\mathtt{x}}$

$$\mathfrak{z}(\alpha + \Delta_{\text{full}}) - \varepsilon_{(\alpha+)} \leq \mathfrak{z}^\circ(\alpha) \leq \mathfrak{z}(\alpha - \Delta_{\text{full}}), \tag{D.12}$$

where

$$\varepsilon_{(\alpha+)} \stackrel{\text{def}}{=} \begin{cases} 0, & \text{if c.d.f. of } L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*) \text{ is continuous in } \mathfrak{z}(\alpha + \Delta_{\text{full}}); \\ \mathtt{C}(p+\mathtt{x})/\sqrt{n} \text{ s.t. (D.44) is fulfilled,} & \text{otherwise.} \end{cases} \tag{D.13}$$

$$I\!P\left(\sqrt{2\{L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*)\}} > \mathfrak{z}(\alpha + \Delta_{\text{full}}) - \varepsilon_{(\alpha+)}\right) \geq \alpha + \Delta_{\text{full}}. \tag{D.14}$$

Indeed, due to Theorem 2.1 and definition (2.2)

$$\begin{aligned} &I\!P^\circ\left(\sqrt{2\Big\{L^\circ(\widetilde{\boldsymbol{\theta}}^\circ) - L^\circ(\widetilde{\boldsymbol{\theta}})\Big\}} > \mathfrak{z}(\alpha - \Delta_{\text{full}})\right) \\ &\qquad \leq I\!P\Big(\sqrt{2\{L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*)\}} > \mathfrak{z}(\alpha - \Delta_{\text{full}})\Big) + \Delta_{\text{full}} \leq \alpha, \end{aligned}$$

therefore, by definition (2.3) $\mathfrak{z}^\circ(\alpha) \leq \mathfrak{z}(\alpha - \Delta_{\text{full}})$. The lower bound in (D.12) is derived similarly. Denote

$$\varepsilon_{(\alpha-)} \stackrel{\text{def}}{=} \begin{cases} 0, & \text{if c.d.f. of } L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*) \text{ is continuous in } \mathfrak{z}(\alpha - \Delta_{\text{full}}); \\ \mathtt{C}(p+\mathtt{x})/\sqrt{n} \text{ s.t. (D.16) is fulfilled,} & \text{otherwise.} \end{cases} \tag{D.15}$$

$$I\!P\left(\sqrt{2\{L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*)\}} > \mathfrak{z}(\alpha - \Delta_{\text{full}}) - \varepsilon_{(\alpha-)}\right) \geq \alpha - \Delta_{\text{full}}. \tag{D.16}$$

Combining (D.16) and Lemma D.1, we obtain with probability $\geq 1 - 12\mathrm{e}^{-\mathtt{x}}$:

$$\begin{aligned} &I\!P\Big(\sqrt{2\{L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*)\}} > \mathfrak{z}^\circ(\alpha)\Big) - \alpha \\ &\quad \geq I\!P\Big(\sqrt{2\{L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*)\}} > \mathfrak{z}(\alpha - \Delta_{\text{full}})\Big) - \alpha \\ &\quad \geq I\!P\Big(\sqrt{2\{L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*)\}} > \mathfrak{z}(\alpha - \Delta_{\text{full}}) - \varepsilon_{(\alpha-)}\Big) - \alpha - \Delta_{\text{LR}} - \varepsilon_{(\alpha-)}/\sqrt{2} \\ &\quad \geq -\Delta_{\text{LR}} - \Delta_{\text{full}} - \varepsilon_{(\alpha-)}/\sqrt{2}. \end{aligned}$$

And similarly for the upper bound

$$
\begin{aligned}
&I\!P\left(\sqrt{2\{L(\widetilde{\boldsymbol{\theta}})-L(\boldsymbol{\theta}^*)\}} > \mathfrak{z}^{\circ}(\alpha)\right) - \alpha \\
&\quad \le I\!P\left(\sqrt{2\{L(\widetilde{\boldsymbol{\theta}})-L(\boldsymbol{\theta}^*)\}} > \mathfrak{z}(\alpha+\Delta_{\mathrm{full}}) - \varepsilon_{(\alpha+)}\right) - \alpha \\
&\quad \le I\!P\left(\sqrt{2\{L(\widetilde{\boldsymbol{\theta}})-L(\boldsymbol{\theta}^*)\}} > \mathfrak{z}(\alpha+\Delta_{\mathrm{full}})\right) - \alpha + \Delta_{\mathrm{LR}} + \varepsilon_{(\alpha+)}/\sqrt{2} \\
&\quad \le \Delta_{\mathrm{full}} + \Delta_{\mathrm{LR}} + \varepsilon_{(\alpha+)}/\sqrt{2}.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\Delta_{\mathfrak{z},\,\mathrm{full}} &\stackrel{\mathrm{def}}{=} \Delta_{\mathrm{full}} + \Delta_{\mathrm{LR}} + \max\left\{\varepsilon_{(\alpha+)}, \varepsilon_{(\alpha-)}\right\}/\sqrt{2} \\
&\le 2.71\bigl(3\delta_n^{1/4} + \breve{\delta}_n^{1/4}\bigr) + \frac{2\sqrt{p}}{3}\left(\delta_{\mathcal{V}}^2(\mathtt{x}) + \delta_{\mathrm{smb}}^2\right) \qquad \text{(D.17)} \\
&\quad + \left\{3\Delta_{\mathrm{W}}(\mathtt{r}_0,\mathtt{x}) + \Delta_{\mathrm{W}}^{\circ}(\mathtt{r}_0,\mathtt{x}) + \Delta_{\xi}^{\circ}(\mathtt{r}_0,\mathtt{x})\right\}/\sqrt{2} \\
&\quad + \mathtt{C}(p+\mathtt{x})/\sqrt{n} \\
&= \mathtt{C}\left\{\frac{(p+\mathtt{x})^3}{n}\right\}^{1/8} + \mathtt{C}\,\frac{p+\mathtt{x}}{\sqrt{n}}\sqrt{\mathtt{x}} + \mathtt{C}\,\frac{p+\mathtt{x}}{\sqrt{n}} \qquad \text{(D.18)} \\
&\quad \text{in the case A.3.1 .}
\end{aligned}
$$

Now let us check the condition $\mathfrak{z}^{\circ}(\alpha) \ge \mathtt{C}\max\{2,\sqrt{p}\} + \mathtt{C}(p+\mathtt{x})/\sqrt{n}$, which comes from the first part of the statement. By Theorems A.4, B.1 and Lemmas A.2, D.2 it holds with probability $\ge 1 - 12\mathrm{e}^{-\mathtt{x}}$:

$$
\begin{aligned}
&I\!P^{\circ}\left(\sqrt{2\{L^{\circ}(\widetilde{\boldsymbol{\theta}}^{\circ}) - L^{\circ}(\widetilde{\boldsymbol{\theta}})\}} > \mathtt{C}\sqrt{p - \sqrt{2\mathtt{x}p}} + \mathtt{C}(p+\mathtt{x})/\sqrt{n}\right) \\
&\qquad \ge 1 - 8\mathrm{e}^{-\mathtt{x}},
\end{aligned}
$$

Taking $1 - 8\mathrm{e}^{-\mathtt{x}} > \alpha$, we have

$$
\mathfrak{z}^{\circ}(\alpha) \ge \mathtt{C}\sqrt{p - \sqrt{2\mathtt{x}p}} + \mathtt{C}2(p+\mathtt{x})/\sqrt{n}.
$$

□

Remark D.1. It is clear from expression (D.11), that the impact of the error term, induced by the Gaussian approximation, is the biggest. The requirement for the ratio $(p+\mathtt{x})^3/n$ to be small is imposed by our Gaussian approximation results (see also Remark B.2 about the multivariate GAR).

Let us introduce for $p=1$ similarly to (B.7) and (D.2)

$$
\delta_{n,\mathrm{B.E.}} \stackrel{\mathrm{def}}{=} \sum_{i=1}^{n} I\!E|\boldsymbol{\xi}_i|^3, \qquad \breve{\delta}_{n,\mathrm{B.E.}} \stackrel{\mathrm{def}}{=} \sum_{i=1}^{n} I\!E^{\circ}|\boldsymbol{\xi}_i^{\circ}(\boldsymbol{\theta}^*)|^3. \qquad \text{(D.19)}
$$

Proof of Theorem 2.3. On a random set of probability $\geq 1 - 12\mathrm{e}^{-\mathtt{x}}$ it holds:

$$
\begin{aligned}
&I\!\!P^{\circ}\left(\sqrt{2\Big\{L^{\circ}(\widetilde{\boldsymbol{\theta}}^{\circ}) - L^{\circ}(\widetilde{\boldsymbol{\theta}})\Big\}} > \mathfrak{z}\right) \\
&\overset{(\text{Th. } A.4)}{\geq} I\!\!P^{\circ}\left(\|\boldsymbol{\xi}^{\circ}(\widetilde{\boldsymbol{\theta}})\| > \mathfrak{z} + \Delta^{\circ}_{\mathrm{W}}(\mathtt{r}_0, \mathtt{x})\right) \\
&\overset{(\text{L. } A.2)}{\geq} I\!\!P^{\circ}\left(\|\boldsymbol{\xi}^{\circ}(\boldsymbol{\theta}^*)\| > \mathfrak{z} + \Delta^{\circ}_{\mathrm{W}}(\mathtt{r}_0, \mathtt{x}) + \Delta^{\circ}_{\xi}(\mathtt{r}_0, \mathtt{x})\right) \\
&\overset{(\text{L. } B.1, \text{Prop. } D.1)}{\geq} I\!\!P\left(\|\boldsymbol{\xi}\| > \mathfrak{z} - \Delta_{\mathrm{W}}(\mathtt{r}_0, \mathtt{x})\right) - \Delta_{\text{B. E., full}} \\
&\overset{(\text{Th. } A.2)}{\geq} I\!\!P\left(\sqrt{2\big\{L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*)\big\}} > \mathfrak{z}\right) - \Delta_{\text{B. E., full}},
\end{aligned}
$$

where the value $\Delta_{\text{B. E., full}}$ comes from the bound (B.10) with $\delta_z := \Delta_{\mathrm{W}}(\mathtt{r}_0, \mathtt{x}) + \Delta^{\circ}_{\mathrm{W}}(\mathtt{r}_0, \mathtt{x}) + \Delta^{\circ}_{\xi}(\mathtt{r}_0, \mathtt{x})$, $C_0 \in [0.4097, 0.560]$ and

$$
\begin{aligned}
\operatorname{Var}^{\circ}\{\boldsymbol{\xi}^{\circ}(\boldsymbol{\theta}^*)\} &\geq \{1 - \delta^2_{\mathcal{V}}(\mathtt{x})\} I\!\!E \operatorname{Var}^{\circ}\{\boldsymbol{\xi}^{\circ}(\boldsymbol{\theta}^*)\} \\
&\geq \frac{3}{4} D_0^{-1} H_0^2 D_0^{-1} \quad \text{for } \delta^2_{\mathcal{V}}(\mathtt{x}) \leq 1/4
\end{aligned}
$$

with probability $\geq 1 - \mathrm{e}^{-\mathtt{x}}$ (due to the bound (D.5)):

$$
\begin{aligned}
\Delta_{\text{B. E., full}} &\overset{\text{def}}{=} 2C_0 \left\{ \frac{\delta_{n,\text{B. E.}}}{(\operatorname{Var}\boldsymbol{\xi})^{3/2}} + \frac{\breve{\delta}_{n,\text{B. E.}}}{(I\!\!E \operatorname{Var}^{\circ}\{\boldsymbol{\xi}^{\circ}(\boldsymbol{\theta}^*)\})^{3/2}} \left(\frac{2}{\sqrt{3}}\right)^3 \right\} \qquad \text{(D.20)} \\
&\quad + \frac{1}{\sqrt{2}} \left\{ \Delta_{\mathrm{W}}(\mathtt{r}_0, \mathtt{x}) + \Delta^{\circ}_{\mathrm{W}}(\mathtt{r}_0, \mathtt{x}) + \Delta^{\circ}_{\xi}(\mathtt{r}_0, \mathtt{x}) \right\} + \frac{2}{3} \left\{ \delta^2_{\mathcal{V}}(\mathtt{x}) + \delta^2_{\xi} \right\} \\
&\quad \leq \mathtt{C} \frac{1 + \mathtt{x}}{\sqrt{n}} \quad \text{in the case A.3.1.}
\end{aligned}
$$

The similar inequalities in the inverse direction finish the proof of the first part. The second part of the statement of the statement is proved using the same arguments as in the proof of the Theorem's 2.1 second part (starting from inequality (D.12)). Applying Lemma B.1 instead of Theorem B.1 in the proof of Lemma D.1 yields the improved error term from the anti-concentration inequality for the likelihood ratio statistics: $\Delta_{\text{B.E., LR}} \overset{\text{def}}{=} 4C_0 \frac{\delta_{n,\text{B.E.}}}{(\operatorname{Var}\boldsymbol{\phi})^{3/2}} + \sqrt{2}\Delta_{\mathrm{W}}(\mathtt{r}_0, \mathtt{x})$. Thus, the total approximating error term is defined as

$$
\begin{aligned}
\Delta_{\text{B. E. }\mathfrak{z}\text{, full}} &\overset{\text{def}}{=} \Delta_{\text{B. E., full}} + \Delta_{\text{B.E., LR}} + \mathtt{C}(1 + \mathtt{x})/\sqrt{n} \qquad \text{(D.21)} \\
&\leq \mathtt{C}(1 + \mathtt{x})/\sqrt{n} \quad \text{in the case A.3.1.}
\end{aligned}
$$

The condition on the quantile $\mathfrak{z}_\alpha$ is implied similarly as in the proof of Theorem 2.2. □

D.1.2 Proof of Theorem 2.4 (large modelling bias)

Lemma D.2 (Lower bound for deviations of a Gaussian quadratic form)**.** *Let* $\boldsymbol{\phi} \sim \mathcal{N}(0, \boldsymbol{I}_p)$ *and* Σ *is any symmetric non-negative definite matrix, then it holds for any* $\mathtt{x} > 0$

$$\mathbb{P}\left(\operatorname{tr}\Sigma - \|\Sigma^{1/2}\boldsymbol{\phi}\|^2 \geq 2\sqrt{\mathtt{x}\operatorname{tr}(\Sigma^2)}\right) \leq \exp(-\mathtt{x}).$$

Proof of Lemma D.2. It is sufficient to consider w.l.o.g. only the case of diagonal matrix Σ, since it can be represented as $\Sigma = U^\top \operatorname{diag}\{a_1, \ldots, a_p\}U$ for an orthogonal matrix U and the eigenvalues $a_1 \geq \cdots \geq a_p$; $U\boldsymbol{\phi} \sim \mathcal{N}(0, \boldsymbol{I}_p)$.

By the exponential Chebyshev inequality it holds for $\mu > 0$, $\Delta > 0$

$$\mathbb{P}\left(\operatorname{tr}\Sigma - \|\Sigma^{1/2}\boldsymbol{\phi}\|^2 \geq \Delta\right) \leq \exp(-\mu\Delta/2)\mathbb{E}\exp\left(\mu\left\{\operatorname{tr}\Sigma - \|\Sigma^{1/2}\boldsymbol{\phi}\|^2\right\}/2\right).$$

$$\log\mathbb{E}\exp\left(\mu\left\{\operatorname{tr}\Sigma - \|\Sigma^{1/2}\boldsymbol{\phi}\|^2\right\}/2\right) \leq \frac{1}{2}\sum_{j=1}^{p}\left\{\mu a_j - \log(1 + a_j\mu)\right\},$$

therefore

$$\begin{aligned}
\mathbb{P}\left(\operatorname{tr}\Sigma - \|\Sigma^{1/2}\boldsymbol{\phi}\|^2 \geq \Delta\right) &\leq \exp\left(-\frac{1}{2}\left[\mu\Delta + \textstyle\sum_{j=1}^{p}\left\{\log(1 + a_j\mu) - \mu a_j\right\}\right]\right) \\
&\leq \exp\left(-\frac{1}{2}\left[\mu\Delta - \mu^2\textstyle\sum_{j=1}^{p} a_j^2/2\right]\right) \\
&\leq \exp\left(-\Delta^2/\left\{4\textstyle\sum_{j=1}^{p} a_j^2\right\}\right).
\end{aligned}$$

If $\mathtt{x} := \Delta^2/\left\{4\sum_{i=1}^{p} a_j^2\right\}$, then $\Delta = 2\sqrt{\mathtt{x}\sum_{j=1}^{p} a_j^2}$. □

Proof of Theorem 2.4. Due to the bound (D.6) it holds for $\mathfrak{z} \geq \max\{2, \sqrt{p}\} + \mathtt{C}(p + \mathtt{x})/\sqrt{n}$ with probability $\geq 1 - 5\mathrm{e}^{-\mathtt{x}}$

$$\begin{aligned}
&\mathbb{P}^\circ\left(\sqrt{2\left\{L^\circ(\widetilde{\boldsymbol{\theta}}^\circ) - L^\circ(\widetilde{\boldsymbol{\theta}})\right\}} > \mathfrak{z}\right) \\
&\qquad \geq \mathbb{P}^\circ\left(\|\boldsymbol{\xi}^\circ(\boldsymbol{\theta}^*)\| > \mathfrak{z} + \Delta_{\mathrm{W}}^\circ(\mathtt{r}_0, \mathtt{x}) + \Delta_{\xi}^\circ(\mathtt{r}_0, \mathtt{x})\right).
\end{aligned}$$

Let us introduce the random vector $\boldsymbol{\xi}_0 \stackrel{\mathrm{def}}{=} (D_0^{-1}H_0^2D_0^{-1})^{1/2}(\operatorname{Var}\boldsymbol{\xi})^{-1/2}\boldsymbol{\xi}$. The bound (D.5) implies with probability $\geq 1 - \mathrm{e}^{-\mathtt{x}}$

$$\operatorname{tr}\left\{\left((\operatorname{Var}\boldsymbol{\xi}_0)^{-1/2}\operatorname{Var}^\circ\left\{\boldsymbol{\xi}^\circ(\boldsymbol{\theta}^*)\right\}(\operatorname{Var}\boldsymbol{\xi}_0)^{-1/2} - \boldsymbol{I}_p\right)^2\right\} \leq p\delta_{\mathcal{V}}^4(\mathtt{x}). \tag{D.22}$$

Applying statement *2.2* of Theorem B.1 to the vectors $\boldsymbol{\xi}^\circ(\boldsymbol{\theta}^*)$ and $\boldsymbol{\xi}_0$, we have with probability $\geq 1 - \mathrm{e}^{-\mathtt{x}}$

$$\begin{aligned}
&\mathbb{P}^\circ\left(\|\boldsymbol{\xi}^\circ(\boldsymbol{\theta}^*)\| > \mathfrak{z} + \Delta_{\mathrm{W}}^\circ(\mathtt{r}_0, \mathtt{x}) + \Delta_{\xi}^\circ(\mathtt{r}_0, \mathtt{x})\right) \\
&\qquad \geq \mathbb{P}\left(\|\boldsymbol{\xi}_0\| > \mathfrak{z} - \Delta_{\mathrm{W}}(\mathtt{r}_0, \mathtt{x})\right) - \Delta_{\mathrm{b,\,full}}
\end{aligned}$$

where

$$\begin{aligned}\Delta_{\mathrm{b,\,full}} \stackrel{\mathrm{def}}{=} 2.71\left(\delta_n^{1/4}+\breve{\delta}_n^{1/4}\right)+\frac{\sqrt{p}}{2}\delta^2_{\mathcal{V}}(\mathtt{x}) \\ + \frac{\Delta_{\mathrm{W}}(\mathtt{r}_0,\mathtt{x})+\Delta^{\circ}_{\mathrm{W}}(\mathtt{r}_0,\mathtt{x})+\Delta^{\circ}_{\xi}(\mathtt{r}_0,\mathtt{x})}{\sqrt{2}}.\end{aligned} \tag{D.23}$$

By the definition of $\boldsymbol{\xi}_0$ it holds $\|\boldsymbol{\xi}_0\| \geq \|\boldsymbol{\xi}\|\|(\operatorname{Var}\boldsymbol{\xi})^{1/2}(D_0^{-1}H_0^2D_0^{-1})^{-1/2}\|^{-1}$. Consider the following matrix

$$\begin{aligned}\widetilde{V}^2 &\stackrel{\mathrm{def}}{=} (D_0^{-1}H_0^2D_0^{-1})^{-1/2}(\operatorname{Var}\boldsymbol{\xi})(D_0^{-1}H_0^2D_0^{-1})^{-1/2} \\ &= (D_0^{-1}H_0^2D_0^{-1})^{1/2}\left(D_0H_0^{-2}V_0^2H_0^{-2}D_0\right)(D_0^{-1}H_0^2D_0^{-1})^{1/2} \\ &\leq (D_0^{-1}H_0^2D_0^{-1})^{1/2}\left(D_0H_0^{-2}D_0\right)(D_0^{-1}H_0^2D_0^{-1})^{1/2} \\ &= \boldsymbol{I}_p,\end{aligned} \tag{D.24}$$

here $V_0^2 \stackrel{\mathrm{def}}{=} \operatorname{Var}\{\nabla_{\boldsymbol{\theta}}L(\boldsymbol{\theta}^*)\}$; the inequality (D.24) holds due to the definitions (1.9), (1.10) and $V_0^2 = H_0^2 - B_0^2 > 0$. Therefore $\|\widetilde{V}^2\| \leq 1$ and $\|\boldsymbol{\xi}_0\| \geq \|\boldsymbol{\xi}\|$. By (D.8)

$$\begin{aligned}&I\!P^{\circ}\left(\sqrt{2\left\{L^{\circ}(\widetilde{\boldsymbol{\theta}}^{\circ})-L^{\circ}(\widetilde{\boldsymbol{\theta}})\right\}} > \mathfrak{z}\right) \\ &\quad\geq I\!P\left(\|\boldsymbol{\xi}\| > \mathfrak{z} - \Delta_{\mathrm{W}}(\mathtt{r}_0,\mathtt{x})\right) - \Delta_{\mathrm{b,\,full}} \\ &\quad\geq I\!P\left(\sqrt{2\left\{L(\widetilde{\boldsymbol{\theta}})-L(\boldsymbol{\theta}^*)\right\}} > \mathfrak{z}\right) - \Delta_{\mathrm{b,\,full}}\end{aligned}$$

with probability $\geq 1 - 12\mathrm{e}^{-\mathtt{x}}$, which finishes the proof of the first part. For the second part let us introduce $\overline{\boldsymbol{\xi}}_0 \sim \mathcal{N}(0, D_0^{-1}H_0^2D_0^{-1})$ s.t. $\operatorname{Var}\overline{\boldsymbol{\xi}}_0 = \operatorname{Var}\boldsymbol{\xi}_0$. Applying statement *2.1* of Theorem B.1 to the vectors $\boldsymbol{\xi}^{\circ}(\boldsymbol{\theta}^*)$ and $\overline{\boldsymbol{\xi}}_0$, using the bound (D.22), we have with probability $\geq 1 - \mathrm{e}^{-\mathtt{x}}$

$$\begin{aligned}&I\!P^{\circ}\left(\|\boldsymbol{\xi}^{\circ}(\boldsymbol{\theta}^*)\| > \mathfrak{z} + \Delta^{\circ}_{\mathrm{W}}(\mathtt{r}_0,\mathtt{x}) + \Delta^{\circ}_{\xi}(\mathtt{r}_0,\mathtt{x})\right) \\ &\quad\geq I\!P\left(\|\overline{\boldsymbol{\xi}}_0\| > \mathfrak{z}\right) - \Delta_{\mathrm{G},1},\end{aligned}$$

where

$$\Delta_{\mathrm{G},1} \stackrel{\mathrm{def}}{=} 2.71\breve{\delta}_n^{1/4} + \frac{\Delta^{\circ}_{\mathrm{W}}(\mathtt{r}_0,\mathtt{x}) + \Delta^{\circ}_{\xi}(\mathtt{r}_0,\mathtt{x})}{\sqrt{2}} + \frac{\sqrt{p}}{2}\delta^2_{\mathcal{V}}(\mathtt{x}). \tag{D.25}$$

By definition (D.1) $\overline{\boldsymbol{\xi}} \sim \mathcal{N}(0, \operatorname{Var}\boldsymbol{\xi})$. Lemma D.2 and Theorem 1.2 by Spokoiny (2012b) imply

$$\begin{aligned}&I\!P\Big(\|\overline{\boldsymbol{\xi}}\| - \|\overline{\boldsymbol{\xi}}_0\| \geq \sqrt{\operatorname{tr}(\operatorname{Var}\boldsymbol{\xi})} - \sqrt{\operatorname{tr}(\operatorname{Var}\overline{\boldsymbol{\xi}}_0)} + \Delta_{\mathtt{qf},1}\Big) \leq 2\mathrm{e}^{-\mathtt{x}}, \\ &I\!P\Big(\|\overline{\boldsymbol{\xi}}\| - \|\overline{\boldsymbol{\xi}}_0\| \leq \sqrt{\operatorname{tr}(\operatorname{Var}\boldsymbol{\xi})} - \sqrt{\operatorname{tr}(\operatorname{Var}\overline{\boldsymbol{\xi}}_0)} - \Delta_{\mathtt{qf},2}\Big) \leq 2\mathrm{e}^{-\mathtt{x}},\end{aligned} \tag{D.26}$$

where

$$\begin{aligned}
\Delta_{\mathtt{qf},1} &\stackrel{\text{def}}{=} \left[4\mathtt{x}\,\mathrm{tr}\{(\mathrm{Var}\,\overline{\boldsymbol{\xi}}_0)^2\}\right]^{1/4} \\
&\quad + \max\left[2\sqrt{2\mathtt{x}\,\mathrm{tr}\{(\mathrm{Var}\,\boldsymbol{\xi})^2\}}, 6\mathtt{x}\|\,\mathrm{Var}\,\boldsymbol{\xi}\|\right]^{1/2}, \\
\Delta_{\mathtt{qf},2} &\stackrel{\text{def}}{=} \left[4\mathtt{x}\,\mathrm{tr}\{(\mathrm{Var}\,\boldsymbol{\xi})^2\}\right]^{1/4} \\
&\quad + \max\left[2\sqrt{2\mathtt{x}\,\mathrm{tr}\{(\mathrm{Var}\,\overline{\boldsymbol{\xi}}_0)^2\}}, 6\mathtt{x}\|\,\mathrm{Var}\,\overline{\boldsymbol{\xi}}_0\|\right]^{1/2}.
\end{aligned} \tag{D.27}$$

By conditions $(\mathcal{I})$, $(\mathcal{I}_B)$

$$\begin{aligned}
\Delta_{\mathtt{qf},1} &\le \left\{\sqrt{4\mathtt{x}p}(\mathfrak{a}^2+\mathfrak{a}_B^2)\right\}^{1/2} + \mathfrak{a}\max\left\{\sqrt{8\mathtt{x}p}, 6\mathtt{x}\right\}^{1/2}, \\
\Delta_{\mathtt{qf},2} &\le \left\{4\mathtt{x}p\mathfrak{a}^4\right\}^{1/4} + \sqrt{\mathfrak{a}^2+\mathfrak{a}_B^2}\max\left\{\sqrt{8\mathtt{x}p}, 6\mathtt{x}\right\}^{1/2}.
\end{aligned} \tag{D.28}$$

Further, it holds on a random set with probability $\ge 1-2\mathrm{e}^{-\mathtt{x}}$

$$\begin{aligned}
&I\!P\left(\|\overline{\boldsymbol{\xi}}_0\| > \mathfrak{z}\right) - \Delta_{\mathrm{G},1} \\
&\quad = I\!P\left(\|\overline{\boldsymbol{\xi}}\| > \mathfrak{z} + \|\overline{\boldsymbol{\xi}}\| - \|\overline{\boldsymbol{\xi}}_0\|\right) - \Delta_{\mathrm{G},1} \\
&\quad \stackrel{(\text{by (D.26)})}{\ge} I\!P\left(\|\overline{\boldsymbol{\xi}}\| > \mathfrak{z} + \sqrt{\mathrm{tr}(\mathrm{Var}\,\boldsymbol{\xi})} - \sqrt{\mathrm{tr}(\mathrm{Var}\,\overline{\boldsymbol{\xi}}_0)} + \Delta_{\mathtt{qf},1}\right) - \Delta_{\mathrm{G},1} \\
&\quad \stackrel{(\text{Th. } B.1)}{\ge} I\!P\left(\|\boldsymbol{\xi}\| > \mathfrak{z} - \Delta_{\mathrm{W}}(\mathtt{r}_0,\mathtt{x}) + \sqrt{\mathrm{tr}(\mathrm{Var}\,\boldsymbol{\xi})} - \sqrt{\mathrm{tr}(\mathrm{Var}\,\overline{\boldsymbol{\xi}}_0)} + \Delta_{\mathtt{qf},1}\right) \\
&\quad - \Delta_{\mathrm{G},1} - \Delta_{\mathrm{G},2} \\
&\quad \stackrel{(\text{Th. } A.2)}{\ge} I\!P\left(\sqrt{2\{L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*)\}} > \mathfrak{z} + \sqrt{\mathrm{tr}(\mathrm{Var}\,\boldsymbol{\xi})} - \sqrt{\mathrm{tr}(\mathrm{Var}\,\overline{\boldsymbol{\xi}}_0)} + \Delta_{\mathtt{qf},1}\right) \\
&\quad - \Delta_{\mathrm{b,\,full}},
\end{aligned}$$

where

$$\begin{aligned}
\Delta_{\mathrm{G},2} &\stackrel{\text{def}}{=} 2.71\delta_n^{1/4} + \frac{\Delta_{\mathrm{W}}(\mathtt{r}_0,\mathtt{x})}{\sqrt{2}}, \\
\Delta_{\mathrm{b,\,full}} &= \Delta_{\mathrm{G},1} + \Delta_{\mathrm{G},2}.
\end{aligned}$$

Hence, we obtain

$$\begin{aligned}
&I\!P^\circ\left(\sqrt{2\left\{L^\circ(\widetilde{\boldsymbol{\theta}}^\circ) - L^\circ(\widetilde{\boldsymbol{\theta}})\right\}} > \mathfrak{z}\right) \\
&\quad \ge I\!P\left(\sqrt{2\{L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*)\}} > \mathfrak{z} + \sqrt{\mathrm{tr}(\mathrm{Var}\,\boldsymbol{\xi})} - \sqrt{\mathrm{tr}(\mathrm{Var}\,\overline{\boldsymbol{\xi}}_0)} + \Delta_{\mathtt{qf},1}\right) \\
&\quad - \Delta_{\mathrm{b,\,full}}.
\end{aligned}$$

By definition (2.2) of $(1-\alpha)$-quantile $\mathfrak{z}(\alpha)$ it holds:

$$\mathfrak{z}(\alpha + \Delta_{\mathrm{b,\,full}}) \le \mathfrak{z}^{\circ}(\alpha) + \sqrt{\mathrm{tr}(\mathrm{Var}\,\boldsymbol{\xi})} - \sqrt{\mathrm{tr}(\mathrm{Var}\,\overline{\boldsymbol{\xi}}_0)} + \mathtt{C}\Delta_{\mathtt{qf},1},$$

and in addition

$$\sqrt{\mathrm{tr}(\mathrm{Var}\,\boldsymbol{\xi})} - \sqrt{\mathrm{tr}(\mathrm{Var}\,\overline{\boldsymbol{\xi}}_0)} \le -\frac{\mathrm{tr}(D_0^{-1}B_0^2D_0^{-1})}{2\sqrt{\mathrm{tr}(D_0^{-1}H_0^2D_0^{-1})}} \le 0.$$

The inverse inequalities are implied with the similar arguments:

$$\begin{aligned}
&I\!P^{\circ}\left(\sqrt{2\bigl\{L^{\circ}(\widetilde{\boldsymbol{\theta}}^{\circ}) - L^{\circ}(\widetilde{\boldsymbol{\theta}})\bigr\}} > \mathfrak{z}\right) \\
&\quad \le I\!P\left(\sqrt{2\bigl\{L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*)\bigr\}} > \mathfrak{z} + \sqrt{\mathrm{tr}(\mathrm{Var}\,\boldsymbol{\xi})} - \sqrt{\mathrm{tr}(\mathrm{Var}\,\overline{\boldsymbol{\xi}}_0)} - \Delta_{\mathtt{qf},2}\right) \\
&\quad + \Delta_{\mathrm{b,\,full}}.
\end{aligned}$$

And

$$\mathfrak{z}(\alpha - \Delta_{\mathrm{b,\,full}}) \ge \mathfrak{z}^{\circ}(\alpha) + \sqrt{\mathrm{tr}(\mathrm{Var}\,\boldsymbol{\xi})} - \sqrt{\mathrm{tr}(\mathrm{Var}\,\overline{\boldsymbol{\xi}}_0)} - \Delta_{\mathtt{qf},2}.$$

□

D.1.3 Proof of Theorem 2.5 (the smoothed version)

Lemma D.3. *For the function $g_{\Delta}(x,z)$ defined in (2.7), all $\Delta_1 \in [0,x]$ and all $C \ge 1$ it holds*

$$g_{\Delta}(x - \Delta_1, z) \ge g_{\Delta}(x, z + \Delta_1 C)$$

Proof of Lemma D.3. By definition (B.11) of $g(x)$

$$\begin{aligned}
\sup\{x \ge 0 : g_{\Delta}(x - \Delta_1, z) = 0\} &= z + \Delta_1, \\
\sup\{x \ge 0 : g_{\Delta}(x, z + \Delta_1 C) = 0\} &= z + \Delta_1 C.
\end{aligned}$$

For $x \ge z + \Delta_1 C$ it holds

$$\begin{aligned}
g_{\Delta}(x - \Delta_1, z) &= g\left(\frac{1}{2\Delta z}\left\{(x-\Delta_1)^2 - z^2\right\}\right) \\
&\ge g\left(\frac{1}{2\Delta(z + \Delta_1 C)}\left\{x^2 - (z + \Delta_1 C)^2\right\}\right) \qquad \text{(D.29)} \\
&= g_{\Delta}(x, z + \Delta_1 C).
\end{aligned}$$

Indeed, the comparison in (D.29) reads as

$$\begin{aligned}(z+\Delta_1 C)(x-\Delta_1+z)(x-\Delta_1-z) \\ \vee\, z(x+z+\Delta_1 C)(x-z-\Delta_1 C).\end{aligned} \tag{D.30}$$

Since $C \geq 1$, $(x-\Delta_1-z) \geq (x-\Delta_1 C-z)$ and it holds for the left side of (D.30):

$$\begin{aligned}(z+\Delta_1 C)(x-\Delta_1+z) &= (zx+z^2+z\Delta_1 C)+\Delta_1(xC-\Delta_1 C-z) \\ &\geq (zx+z^2+z\Delta_1 C),\end{aligned}$$

which is equal to the multiplier $z(x+\Delta_1 C+z)$ on the right side of (D.30). □

Proposition D.2 (Smooth analog of Proposition D.1)**.** *If conditions* **(SmB)** *and* **(SD_1)** *are fulfilled, then it holds for all* $0 < \Delta \leq 0.22$ *and for all* $z, \overline{z} > 2$ *s.t.* $|z-\overline{z}| \leq \delta_z$ *for some* $\delta_z \in [0,1]$ *with probability* $\geq 1-\mathrm{e}^{-\mathtt{x}}$*:*

$$\begin{aligned}&\Big|I\!\!E g_\Delta\left(\|\boldsymbol{\xi}\|, z\right) - I\!\!E^\circ g_\Delta\left(\|\boldsymbol{\xi}^\circ(\boldsymbol{\theta}^*)\|, \overline{z}\right)\Big| \\ &\quad\leq \frac{16}{\Delta^3}\left(\delta_n + \breve{\delta}_n\right) + 2\sqrt{p}\frac{\delta_z}{z} + \sqrt{p}\frac{\delta_z^2}{z^2} + \sqrt{p}\frac{\delta^2_{\mathcal{V}}(\mathtt{x}) + \delta^2_{\mathrm{smb}}}{1-\delta^2_{\mathcal{V}}(\mathtt{x})} \\ &\quad\leq \frac{16}{\Delta^3}\left(\delta_n + \breve{\delta}_n\right) + \sqrt{5}\delta_z + \frac{4\sqrt{p}}{3}\left\{\delta^2_{\mathcal{V}}(\mathtt{x}) + \delta^2_{\mathrm{smb}}\right\} \\ &\quad\text{for } \overline{z} \geq \sqrt{p},\ \delta^2_{\mathcal{V}}(\mathtt{x}) \leq 1/4.\end{aligned} \tag{D.31}$$

Proof of Proposition D.2. The conditions of Theorem B.2 are fulfilled with the value $\delta_\Sigma = \sqrt{p}\left\{\delta^2_{\mathcal{V}}(\mathtt{x}) + \delta^2_{\mathrm{smb}}\right\} / \left\{1-\delta^2_{\mathcal{V}}(\mathtt{x})\right\}$ due to the proof of Proposition D.1. □

Proof of Theorem 2.5. The following holds on a random set of probability $\geq 1-12\mathrm{e}^{-\mathtt{x}}$:

$$\begin{aligned}&I\!\!E^\circ g_\Delta\left(\sqrt{2\Big\{L^\circ(\widetilde{\boldsymbol{\theta}}^\circ) - L^\circ(\widetilde{\boldsymbol{\theta}})\Big\}}, \mathfrak{z}\right) \\ &\quad\overset{(\text{Th. } A.4)}{\geq} I\!\!E^\circ g_\Delta\left(\|\boldsymbol{\xi}^\circ(\widetilde{\boldsymbol{\theta}})\| - \Delta^\circ_{\mathrm{W}}(\mathtt{r}_0,\mathtt{x}), \mathfrak{z}\right) \\ &\quad\overset{(\text{L. } A.2)}{\geq} I\!\!E^\circ g_\Delta\left(\|\boldsymbol{\xi}^\circ(\boldsymbol{\theta}^*)\| - \Delta^\circ_{\mathrm{W}}(\mathtt{r}_0,\mathtt{x}) - \Delta^\circ_{\xi}(\mathtt{r}_0,\mathtt{x}), \mathfrak{z}\right) \\ &\quad\overset{(\text{L. } D.3)}{\geq} I\!\!E^\circ g_\Delta\left(\|\boldsymbol{\xi}^\circ(\boldsymbol{\theta}^*)\|, \mathfrak{z} + \Delta^\circ_{\mathrm{W}}(\mathtt{r}_0,\mathtt{x}) + \Delta^\circ_{\xi}(\mathtt{r}_0,\mathtt{x})\right) \\ &\quad\overset{(\text{Prop. } D.2)}{\geq} I\!\!E g_\Delta\left(\|\boldsymbol{\xi}\|, \mathfrak{z} - \Delta_{\mathrm{W}}(\mathtt{r}_0,\mathtt{x})\right) - \Delta_{\mathrm{sm}} \\ &\quad\overset{(\text{Th. } A.2,\ \text{L. } D.3)}{\geq} I\!\!E g_\Delta\left(\sqrt{2\Big\{L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*)\Big\}}, \mathfrak{z}\right) - \Delta_{\mathrm{sm}},\end{aligned}$$

where the term Δ_{sm} comes from (D.31) with $\delta_z := \Delta_{\mathrm{W}}(\mathtt{r}_0,\mathtt{x}) + \Delta^\circ_{\mathrm{W}}(\mathtt{r}_0,\mathtt{x}) + \Delta^\circ_{\xi}(\mathtt{r}_0,\mathtt{x})$:

$$\begin{aligned}\Delta_{\mathrm{sm}} \overset{\mathrm{def}}{=} &\frac{16}{\Delta^3}\left(\delta_n + \breve{\delta}_n\right) + \frac{4\sqrt{p}}{3}\left\{\delta^2_{\mathcal{V}}(\mathtt{x}) + \delta^2_{\mathrm{smb}}\right\} \\ &+ \sqrt{5}\left\{\Delta_{\mathrm{W}}(\mathtt{r}_0,\mathtt{x}) + \Delta^\circ_{\mathrm{W}}(\mathtt{r}_0,\mathtt{x}) + \Delta^\circ_{\xi}(\mathtt{r}_0,\mathtt{x})\right\}.\end{aligned} \tag{D.32}$$

By the similar inequalities in the inverse direction we get the statement proved. Due to the arguments in the end of the proof of Theorem 2.1 it holds in the case A.3.1

$$\Delta_{\mathrm{sm}} = \mathtt{C}\frac{1}{\Delta^3}\Big\{\frac{(p+\mathtt{x})^3}{n}\Big\}^{1/2} + \mathtt{C}\,\frac{p+\mathtt{x}}{\sqrt{n}}\sqrt{\mathtt{x}} + \mathtt{C}\,\frac{p+\mathtt{x}}{\sqrt{n}}. \tag{D.33}$$

□

D.1.4 Bernstein matrix inequality

Consider the following symmetric $p \times p$ $I\!P$-random matrix and its expected value:

$$\mathcal{V}^2(\boldsymbol{\theta}^*) \stackrel{\mathrm{def}}{=} \mathrm{Var}^{\circ}(\nabla_{\boldsymbol{\theta}} L^{\circ}(\boldsymbol{\theta}^*)) = \sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}}\ell_i(\boldsymbol{\theta}^*)\nabla_{\boldsymbol{\theta}}\ell_i(\boldsymbol{\theta}^*)^\top,$$
$$H_0^2 \stackrel{\mathrm{def}}{=} I\!E\mathcal{V}^2(\boldsymbol{\theta}^*) = \sum_{i=1}^{n} I\!E\left[\nabla_{\boldsymbol{\theta}}\ell_i(\boldsymbol{\theta}^*)\nabla_{\boldsymbol{\theta}}\ell_i(\boldsymbol{\theta}^*)^\top\right].$$

Matrix $\mathcal{V}^2(\boldsymbol{\theta}^*)$ equals to a sum of the independent random matrices $\nabla_{\boldsymbol{\theta}}\ell_i(\boldsymbol{\theta}^*)\nabla_{\boldsymbol{\theta}}\ell_i(\boldsymbol{\theta}^*)^\top$. Assuming the condition $\boldsymbol{(SD_1)}$ to be fulfilled, we can refer to the result by Tropp (2012) in order to get the concentration bound below. Let us previously introduce some notations.

$$v_i^2(\boldsymbol{\theta}) \stackrel{\mathrm{def}}{=} H_0^{-1}\left\{\nabla_{\boldsymbol{\theta}}\ell_i(\boldsymbol{\theta})\nabla_{\boldsymbol{\theta}}\ell_i(\boldsymbol{\theta})^\top - I\!E\left[\nabla_{\boldsymbol{\theta}}\ell_i(\boldsymbol{\theta})\nabla_{\boldsymbol{\theta}}\ell_i(\boldsymbol{\theta})^\top\right]\right\}H_0^{-1},$$

then

$$H_0^{-1}\mathcal{V}^2(\boldsymbol{\theta}^*)H_0^{-1} - \boldsymbol{I}_p = \sum\nolimits_{i=1}^{n} v_i^2(\boldsymbol{\theta}^*). \tag{D.34}$$

Define also

$$\varkappa_v^2 \stackrel{\mathrm{def}}{=} \Big\|\sum\nolimits_{i=1}^{n} I\!E v_i^4(\boldsymbol{\theta}^*)\Big\|.$$

Theorem D.1 (Bernstein inequality for $\mathcal{V}^2(\boldsymbol{\theta}^*)$)**.** *Let the condition* $\boldsymbol{(SD_1)}$ *be fulfilled, then it holds with probability* $\geq 1 - \mathrm{e}^{-\mathtt{x}}$ *:*

$$\|H_0^{-1}\mathcal{V}^2(\boldsymbol{\theta}^*)H_0^{-1} - \boldsymbol{I}_p\| \leq \delta_{\mathcal{V}}^2(\mathtt{x}),$$

where the error term is defined as

$$\delta_{\mathcal{V}}^2(\mathtt{x}) \stackrel{\mathrm{def}}{=} \sqrt{2\varkappa_v^2\left\{\log(p) + \mathtt{x}\right\}} + \frac{2}{3}\delta_v^2\left\{\log(p) + \mathtt{x}\right\} \tag{D.35}$$

and is proportional to $\sqrt{\{\log(p) + \mathtt{x}\}/n}$ *in the case A.3.1.*

Proof. Due to Theorem 1.4 by Tropp (2012):

$$I\!P\big(\|H_0^{-1}\mathcal{V}^2(\boldsymbol{\theta}^*)H_0^{-1} - \boldsymbol{I}_p\| \geq t\big) \leq p \exp\left(\frac{-t^2}{2\varkappa_v^2 + 2\delta_v^2 t/3}\right).$$

For

$$\mathtt{x} = \frac{t^2}{2\varkappa_v^2 + 2\delta_v^2 t/3} - \log(p)$$

it holds:

$$I\!P\left(\|H_0^{-1}\mathcal{V}^2(\boldsymbol{\theta}^*)H_0^{-1} - \boldsymbol{I}_p\| \geq \delta_{\mathcal{V}}^2(\mathtt{x})\right) \leq \mathrm{e}^{-\mathtt{x}}.$$

□

D.2 Proofs for Chapter 3

Before proving the statements from Section 3.2.2 we formulate below the Bernstein matrix inequality, which is necessary for the further proofs.

D.2.1 Bernstein matrix inequality

Here we restate the Theorem 1.4 by Tropp (2012) for the random $p_{\mathrm{sum}} \times p_{\mathrm{sum}}$ matrix $\widehat{\mathcal{V}}^2 \stackrel{\mathrm{def}}{=} \mathrm{Var}^{\circ}\big(\nabla_{\boldsymbol{\theta}} L_1^{\circ}(\boldsymbol{\theta}_1^*)^\top, \ldots, \nabla_{\boldsymbol{\theta}} L_K^{\circ}(\boldsymbol{\theta}_K^*)^\top\big)^\top$ from the bootstrap world. Matrix $\widehat{\mathcal{V}}^2$ equals to the sum of independent matrices $\mathrm{Var}^{\circ}\big(\nabla_{\boldsymbol{\theta}} \ell_{i,1}(\boldsymbol{\theta}_1^*)^\top u_i, \ldots, \nabla_{\boldsymbol{\theta}} \ell_{i,K}(\boldsymbol{\theta}_K^*)^\top u_i\big)^\top$. Let us denote

$$\begin{aligned}
\boldsymbol{g}_i &\stackrel{\mathrm{def}}{=} \left(\nabla_{\boldsymbol{\theta}} \ell_{i,1}(\boldsymbol{\theta}_1^*)^\top, \ldots, \nabla_{\boldsymbol{\theta}} \ell_{i,K}(\boldsymbol{\theta}_K^*)^\top\right)^\top \in I\!R^{p_{\mathrm{sum}}}, \\
\widehat{H}^2 &\stackrel{\mathrm{def}}{=} \sum\nolimits_{i=1}^n I\!E\left\{\boldsymbol{g}_i \boldsymbol{g}_i^\top\right\}, \\
\widehat{v}_i &\stackrel{\mathrm{def}}{=} \widehat{H}^{-1}\left\{\boldsymbol{g}_i \boldsymbol{g}_i^\top - I\!E\left[\boldsymbol{g}_i \boldsymbol{g}_i^\top\right]\right\}\widehat{H}^{-1},
\end{aligned}$$

then

$$\widehat{H}^2 = I\!E\widehat{\mathcal{V}}^2, \qquad \sum\nolimits_{i=1}^n \widehat{v}_i^2 = \widehat{H}^{-1}\widehat{\mathcal{V}}^2\widehat{H}^{-1} - \boldsymbol{I}_{p_{\mathrm{sum}}}.$$

Define also the deterministic scalar value

$$\widehat{\varkappa}_v^2 \stackrel{\mathrm{def}}{=} \Big\|\sum\nolimits_{i=1}^n I\!E\widehat{v}_i^4\Big\|.$$

Theorem D.2 (Bernstein inequality for $\widehat{\mathcal{V}}^2$)**.** *Let the condition* $\widehat{(\boldsymbol{SD}_1)}$ *be fulfilled, then it holds with probability* $\geq 1 - \mathrm{e}^{-\mathtt{x}}$ *:*

$$\|\widehat{H}^{-1}\widehat{\mathcal{V}}^2\widehat{H}^{-1} - \boldsymbol{I}_{p_{\mathrm{sum}}}\| \leq \delta_{\widehat{\mathcal{V}}}^2(\mathtt{x}),$$

where the error term is defined as

$$\delta^2_{\widehat{\mathcal{V}}}(\mathtt{x}) \stackrel{\text{def}}{=} \sqrt{2\widehat{\varkappa}^2_v \{\log(p_{\text{sum}}) + \mathtt{x}\}} + \frac{2}{3}\delta^2_{v^*} \{\log(p_{\text{sum}}) + \mathtt{x}\} \tag{D.36}$$

and is proportional to $\sqrt{\{\log(p_{\text{sum}}) + \mathtt{x}\}/n}$ *in the case A.3.1.*

The statement of Theorem D.2 follows straightforwardly from Theorem 1.4 by Tropp (2012), see also Theorem D.1 in Section D.1.4 above.

D.2.2 Proof of Theorem 3.1

Lemma D.4 (Closeness of $\mathcal{L}\left(\|\boldsymbol{\xi}_1\|, \ldots, \|\boldsymbol{\xi}_K\|\right)$ and $\mathcal{L}^\circ\left(\|\boldsymbol{\xi}^\circ_1\|, \ldots, \|\boldsymbol{\xi}^\circ_K\|\right)$). *If the conditions* $(\boldsymbol{ED_{0,k}})$, $(\boldsymbol{\mathcal{I}_k})$, $(\widehat{\mathbf{SmB}})$, $(\boldsymbol{\mathcal{I}_{B,k}})$, $(\widehat{\boldsymbol{SD}_1})$ *and* $(\boldsymbol{Eb})$ *are fulfilled, then it holds with probability* $\geq 1 - 6\mathrm{e}^{-\mathtt{x}}$ *for all* $\delta_{z_k} \geq 0$ *and* $z_k \geq \sqrt{p_k} + \Delta_\varepsilon$ *s.t.* $\mathtt{C} \max\limits_{1\leq k\leq K}\{n^{-1/2}, \delta_{z_k}\} \leq \Delta_\varepsilon \leq \mathtt{C} \min\limits_{1\leq k\leq K}\{1/z_k\}$ *(*Δ_ε *is given in* (C.3)*):*

$$\begin{aligned}
I\!\!P\left(\bigcup\nolimits_{k=1}^K \{\|\boldsymbol{\xi}_k\| > z_k\}\right) - I\!\!P^\circ\left(\bigcup\nolimits_{k=1}^K \{\|\boldsymbol{\xi}^\circ_k\| > z_k - \delta_{z_k}\}\right) &\geq -\Delta_{\ell_2}, \\
I\!\!P\left(\bigcup\nolimits_{k=1}^K \{\|\boldsymbol{\xi}_k\| > z_k\}\right) - I\!\!P^\circ\left(\bigcup\nolimits_{k=1}^K \{\|\boldsymbol{\xi}^\circ_k\| > z_k + \delta_{z_k}\}\right) &\leq \Delta_{\ell_2}.
\end{aligned}$$

for the deterministic nonnegative value

$$\Delta_{\ell_2} \leq 25\mathtt{C}\left(\frac{p^3_{\max}}{n}\right)^{1/8} \log^{9/8}(K)\log^{3/8}(np_{\text{sum}})\left\{\left(\widehat{\mathfrak{a}}^2 + \widehat{\mathfrak{a}}^2_B\right)\left(1 + \delta^2_{\widehat{\mathcal{V}}}(\mathtt{x})\right)\right\}^{3/8}.$$

A more explicit bound on Δ_{ℓ_2} *is given in Proposition C.1, see also Remark C.1.*

Proof of Lemma D.4. The statement follows from Proposition C.1 and Theorem D.2. Let us take $\boldsymbol{\phi}_k := \boldsymbol{\xi}_k$ and $\boldsymbol{\psi}_k := \boldsymbol{\xi}^\circ_k$. Define similarly to Φ in (C.5)

$$\Xi \stackrel{\text{def}}{=} \left(\boldsymbol{\xi}_1^\top, \ldots, \boldsymbol{\xi}_K^\top\right)^\top \quad \Xi^\circ \stackrel{\text{def}}{=} \left(\boldsymbol{\xi}_1^{\circ\top}, \ldots, \boldsymbol{\xi}_K^{\circ\top}\right)^\top. \tag{D.37}$$

Condition (C.4) rewrites for (D.37) as

$$\|\operatorname{Var}\Xi - \operatorname{Var}^\circ\Xi^\circ\|_{\max} \leq \delta^2_\Sigma$$

for some $\delta^2_\Sigma \geq 0$. Denote

$$\begin{aligned}
\widehat{D}^2 &\stackrel{\text{def}}{=} \operatorname{diag}\left\{D^2_1, \ldots, D^2_K\right\}, \\
\widehat{V}^2 &\stackrel{\text{def}}{=} \operatorname{Var}\left(\nabla_{\boldsymbol{\theta}} L_1(\boldsymbol{\theta}^*_1)^\top, \ldots, \nabla_{\boldsymbol{\theta}} L_K(\boldsymbol{\theta}^*_K)^\top\right)^\top.
\end{aligned}$$

$\widehat{D}^2$ is a block-diagonal matrix and $\widehat{V}^2$ is a block matrix. Both of them are symmetric, positive definite and have the dimension $p_{\mathrm{sum}} \times p_{\mathrm{sum}}$. Let also

$$\begin{aligned}
\widehat{\mathcal{V}}^2 &\stackrel{\mathrm{def}}{=} \mathrm{Var}^{\circ}\left(\nabla_{\boldsymbol{\theta}} L_1^{\circ}(\boldsymbol{\theta}_1^*)^\top, \ldots, \nabla_{\boldsymbol{\theta}} L_K^{\circ}(\boldsymbol{\theta}_K^*)^\top\right)^\top, \\
\boldsymbol{g}_i &\stackrel{\mathrm{def}}{=} \left(\nabla_{\boldsymbol{\theta}} \ell_{i,1}(\boldsymbol{\theta}_1^*)^\top, \ldots, \nabla_{\boldsymbol{\theta}} \ell_{i,K}(\boldsymbol{\theta}_K^*)^\top\right)^\top \in I\!R^{p_{\mathrm{sum}}}, \\
\widehat{H}^2 &\stackrel{\mathrm{def}}{=} \sum\nolimits_{i=1}^n I\!E\left\{\boldsymbol{g}_i \boldsymbol{g}_i^\top\right\}, \quad \widehat{B}^2 \stackrel{\mathrm{def}}{=} \sum\nolimits_{i=1}^n I\!E\left\{\boldsymbol{g}_i\right\} I\!E\left\{\boldsymbol{g}_i\right\}^\top.
\end{aligned}$$

It holds

$$\begin{gathered}
\mathrm{Var}\, \Xi = \widehat{D}^{-1} \widehat{V}^2 \widehat{D}^{-1}, \quad \mathrm{Var}^{\circ}\, \Xi^{\circ} = \widehat{D}^{-1} \widehat{\mathcal{V}}^2 \widehat{D}^{-1}, \\
\widehat{H}^2 = I\!E \widehat{\mathcal{V}}^2, \quad \widehat{V}^2 = \widehat{H}^2 - \widehat{B}^2.
\end{gathered}$$

Therefore

$$\begin{aligned}
\|\mathrm{Var}\, \Xi - \mathrm{Var}^{\circ}\, \Xi^{\circ}\|_{\max} &= \left\|\widehat{D}^{-1}\big(\widehat{V}^2 - \widehat{\mathcal{V}}^2\big)\widehat{D}^{-1}\right\|_{\max} \\
&\le \left\|\widehat{D}^{-1}\big(\widehat{H}^2 - \widehat{\mathcal{V}}^2\big)\widehat{D}^{-1}\right\|_{\max} + \left\|\widehat{D}^{-1}\widehat{B}^2\widehat{D}^{-1}\right\|_{\max} \\
&\le \delta^2_{\widehat{\mathcal{V}}}(\mathtt{x})\left\|\widehat{D}^{-1}\widehat{H}^2\widehat{D}^{-1}\right\| + \left\|\widehat{D}^{-1}\widehat{B}^2\widehat{D}^{-1}\right\| \qquad \text{(D.38)} \\
&\le \left\{\delta^2_{\widehat{\mathcal{V}}}(\mathtt{x}) + \widehat{\delta}^2_{\mathrm{smb}}\right\}(\widehat{\mathfrak{a}}^2 + \widehat{\mathfrak{a}}^2_B) =: \delta^2_{\Sigma}. \qquad \text{(D.39)}
\end{aligned}$$

Here inequality (D.38) follows from the matrix Bernstein inequality by Tropp (2012) (see Section D.2.1). Inequality (D.39) is implied by conditions $(\boldsymbol{\mathcal{I}_{B,k}})$ and $(\widehat{\mathbf{SmB}})$, and Cauchy-Schwarz inequality.

Condition (C1) of Proposition C.1 is fulfilled for the vectors $\boldsymbol{\xi}_{i,k}$ and $\boldsymbol{\xi}^{\circ}_{i,k}$ due to conditions $(\boldsymbol{ED_{0,k}})$, $(\boldsymbol{\mathcal{I}_k})$ and $(\widehat{\boldsymbol{SD}_1})$, $(\boldsymbol{Eb})$, $(\widehat{\mathbf{SmB}})$, $(\boldsymbol{\mathcal{I}_{B,k}})$ for $\mathfrak{c}_\phi := \widehat{\mathfrak{a}}$ and $\mathfrak{c}^2_\psi := \big(\widehat{\mathfrak{a}}^2 + \widehat{\mathfrak{a}}^2_B\big)\left\{\delta^2_{v^*} + \max_{1\le i\le n} \|\widehat{H}^{-1} I\!E\left[\boldsymbol{g}_i \boldsymbol{g}_i^\top\right]\widehat{H}^{-1}\|^2\right\}$. □

Proof of Theorem 3.1. Let us denote $\mathtt{x}_2 \stackrel{\mathrm{def}}{=} \mathtt{x} + \log(K)$. It holds with probability $\ge 1 - 12\mathrm{e}^{-\mathtt{x}}$

$$\begin{aligned}
&I\!P^{\circ}\left(\bigcup\nolimits_{k=1}^K \left\{\sqrt{2L_k^{\circ}(\widetilde{\boldsymbol{\theta}}_k^{\circ}) - 2L_k^{\circ}(\widetilde{\boldsymbol{\theta}}_k)} > z_k\right\}\right) \\
&\stackrel{\text{L. A.7}}{\ge} I\!P^{\circ}\left(\bigcup\nolimits_{k=1}^K \left\{\|\boldsymbol{\xi}_k^{\circ}(\widetilde{\boldsymbol{\theta}}_k)\| \ge z_k + \Delta^{\circ}_{\mathrm{W},k}(\mathtt{r}_{0,k}, \mathtt{x}_2)\right\}\right) \\
&\stackrel{\text{L. A.8}}{\ge} I\!P^{\circ}\left(\bigcup\nolimits_{k=1}^K \left\{\|\boldsymbol{\xi}_k^{\circ}(\boldsymbol{\theta}_k^*)\| > z_k + \Delta^{\circ}_{\mathrm{W},k}(\mathtt{r}_{0,k}, \mathtt{x}_2) + \Delta^{\circ}_{\xi,k}(\mathtt{r}_{0,k}, \mathtt{x}_2)\right\}\right) \\
&\stackrel{\text{L. D.4}}{\ge} I\!P\left(\bigcup\nolimits_{k=1}^K \left\{\|\boldsymbol{\xi}_k\| > z_k - \Delta_{\mathrm{W},k}(\mathtt{r}_{0,k}, \mathtt{x}_2)\right\}\right) - \Delta_{\mathrm{total}} \\
&\stackrel{\text{L. A.7}}{\ge} I\!P\left(\bigcup\nolimits_{k=1}^K \left\{\sqrt{2L_k(\widetilde{\boldsymbol{\theta}}_k) - 2L_k(\boldsymbol{\theta}_k^*)} > z_k\right\}\right) - \Delta_{\mathrm{total}},
\end{aligned}$$

for

$$\Delta_{\text{total}} \stackrel{\text{def}}{=} \Delta_{\ell_2}, \tag{D.40}$$

$$\begin{aligned}\delta_{z_k} &:= \Delta_{\mathrm{w},k}(\mathtt{r}_{0,k}, \mathtt{x}+\log(K)) + \Delta^{\circ}_{\mathrm{w},k}(\mathtt{r}_{0,k}, \mathtt{x}+\log(K)) \qquad \text{(D.41)}\\ &\quad + \Delta^{\circ}_{\xi,k}(\mathtt{r}_{0,k}, \mathtt{x}+\log(K))\\ &\le \mathtt{C}\frac{p_k+\mathtt{x}+\log(K)}{\sqrt{n}}\sqrt{\mathtt{x}+\log(K)} \ \text{ in the case A.3.1.} \qquad \text{(D.42)}\end{aligned}$$

Definition of Δ_{ℓ_2} is given in Proposition C.1, see also Remark C.1. The bound from Lemma D.4 says:

$$\Delta_{\ell_2} \le 25\mathtt{C}\left(\frac{p^3_{\max}}{n}\right)^{1/8} \log^{9/8}(K)\log^{3/8}(np_{\text{sum}})\left\{\left(\widehat{\mathfrak{a}}^2+\widehat{\mathfrak{a}}^2_B\right)\left(1+\delta^2_{\widehat{\gamma}}(\mathtt{x})\right)\right\}^{3/8}.$$

For δ_{z_k} bounded as in (D.42) the conditions $\mathtt{C}\max\limits_{1\le k\le K}\{n^{-1/2}, \delta_{z_k}\} \le \Delta_\varepsilon \le \mathtt{C}\min\limits_{1\le k\le K}\{1/z_k\}$ are fulfilled. □

D.2.3 Proof of Theorem 3.2

Proof of Theorem 3.2. For the pointwise quantile functions $\mathfrak{z}_k(\alpha)$ and $\mathfrak{z}^{\circ}_k(\alpha)$ it holds for each $k=1,\ldots,K$ with dominating probability:

$$\begin{aligned}\mathfrak{z}^{\circ}_k(\alpha+\Delta_{\text{full},k}) &\le \mathfrak{z}_k(\alpha),\\ \mathfrak{z}^{\circ}_k(\alpha) &\ge \mathfrak{z}_k(\alpha+\Delta_{\text{full }k}) - \varepsilon_k\end{aligned} \tag{D.43}$$

here $\Delta_{\text{full},k} \le \left\{(p_k+\mathtt{x})^3/\sqrt{n}\right\}^{1/8}$, it comes from Theorem 2.1, and $\varepsilon_k \le \mathtt{C}(p_k+\mathtt{x})/\sqrt{n}$,

$$\varepsilon_k \stackrel{\text{def}}{=} \begin{cases} 0, & \text{if c.d.f. of } L_k(\widetilde{\boldsymbol{\theta}}_k) - L_k(\boldsymbol{\theta}^*_k) \text{ is continuous in } \mathfrak{z}_k(\alpha+\Delta_{\text{full},k});\\ \mathtt{C}(p_k+\mathtt{x})/\sqrt{n} \ \text{ s.t. (D.44) is fulfilled,} & \text{otherwise.}\end{cases}$$

$$I\!P\left(\sqrt{2\left\{L_k(\widetilde{\boldsymbol{\theta}}_k) - L_k(\boldsymbol{\theta}^*_k)\right\}} > \mathfrak{z}_k(\alpha+\Delta_{\text{full},k}) - \varepsilon_k\right) \ge \alpha+\Delta_{\text{full},k}. \tag{D.44}$$

Indeed, due to Theorem 2.1 and definition (1.15)

$$\begin{aligned}&I\!P^{\circ}\left(\sqrt{2\left\{L^{\circ}_k(\widetilde{\boldsymbol{\theta}}^{\circ}_k) - L^{\circ}_k(\widetilde{\boldsymbol{\theta}}_k)\right\}} > \mathfrak{z}_k(\alpha)\right)\\ &\qquad \le I\!P\left(\sqrt{2\left\{L_k(\widetilde{\boldsymbol{\theta}}_k) - L_k(\boldsymbol{\theta}^*_k)\right\}} > \mathfrak{z}_k(\alpha)\right) + \Delta_{\text{full},k} \le \alpha+\Delta_{\text{full},k},\end{aligned}$$

therefore, by definition (3.3) $\mathfrak{z}^{\circ}_k(\alpha+\Delta_{\text{full},k}) \le \mathfrak{z}_k(\alpha)$. The lower bound is derived similarly.

If there exist the inverse functions $\mathfrak{c}^{-1}(\cdot)$ and $\mathfrak{c}^{\circ\,-1}(\cdot)$, then it holds for $\beta \in (0,1)$:

$$\begin{aligned} I\!P\left(\bigcup\nolimits_{k=1}^{K}\left\{\sqrt{2L_k(\widetilde{\boldsymbol{\theta}}_k) - 2L_k(\boldsymbol{\theta}_k^*)} \geq \mathfrak{z}_k(\beta)\right\}\right) &\leq \mathfrak{c}^{-1}(\beta), \\ I\!P^{\circ}\left(\bigcup\nolimits_{k=1}^{K}\left\{\sqrt{2L_k^{\circ}(\widetilde{\boldsymbol{\theta}}_k^{\circ}) - 2L_k^{\circ}(\widetilde{\boldsymbol{\theta}}_k)} \geq \mathfrak{z}_k^{\circ}(\beta)\right\}\right) &\leq \mathfrak{c}^{\circ\,-1}(\beta). \end{aligned} \tag{D.45}$$

Therefore, it holds

$$\begin{aligned} &\mathfrak{c}^{\circ\,-1}(\beta + \Delta_{\mathrm{full,\,max}}) \\ &\geq I\!P^{\circ}\left(\bigcup\nolimits_{k=1}^{K}\left\{\sqrt{2L_k^{\circ}(\widetilde{\boldsymbol{\theta}}_k^{\circ}) - 2L_k^{\circ}(\widetilde{\boldsymbol{\theta}}_k)} \geq \mathfrak{z}_k^{\circ}(\beta + \Delta_{\mathrm{full},\,k})\right\}\right) \\ &\overset{\text{by (D.43)}}{\geq} I\!P^{\circ}\left(\bigcup\nolimits_{k=1}^{K}\left\{\sqrt{2L_k^{\circ}(\widetilde{\boldsymbol{\theta}}_k^{\circ}) - 2L_k^{\circ}(\widetilde{\boldsymbol{\theta}}_k)} \geq \mathfrak{z}_k(\beta)\right\}\right) \\ &\overset{\text{by Th. 3.1}}{\geq} I\!P\left(\bigcup\nolimits_{k=1}^{K}\left\{\sqrt{2L_k(\widetilde{\boldsymbol{\theta}}_k) - 2L_k(\boldsymbol{\theta}_k^*)} \geq \mathfrak{z}_k(\beta)\right\}\right) - \Delta_{\mathrm{total}} \\ &\overset{\text{by L. D.5 and (D.45)}}{\geq} \mathfrak{c}^{-1}(\beta) - \Delta_{\mathrm{total}} - \Delta_{\mathrm{ac,LR}}, \end{aligned}$$

here $\Delta_{\mathrm{ac,LR}} \leq \Delta_{\mathrm{total}}$ (by Lemma D.5) and

$$\begin{aligned} \Delta_{\mathrm{full,\,max}} &\overset{\mathrm{def}}{=} \max_{1 \leq k \leq K} \Delta_{\mathrm{full},\,k} \\ &\leq \mathtt{C}\{(p_{\max} + \mathtt{x})^3/n\}^{1/8} \text{ in the case A.3.1.} \end{aligned} \tag{D.46}$$

Thus

$$\begin{aligned} \mathfrak{c}^{\circ\,-1}(\beta + \Delta_{\mathrm{full,\,max}}) &\geq \mathfrak{c}^{-1}(\beta) - \Delta_{\mathrm{total}} - \Delta_{\mathrm{ac,LR}}, \\ \mathfrak{c}^{\circ}(\alpha) &\leq \mathfrak{c}(\alpha + \Delta_{\mathrm{total}} + \Delta_{\mathrm{ac,LR}}) + \Delta_{\mathrm{full,\,max}}. \end{aligned} \tag{D.47}$$

Hence it holds

$$\begin{aligned} &I\!P\left(\bigcup\nolimits_{k=1}^{K}\left\{\sqrt{2L_k(\widetilde{\boldsymbol{\theta}}_k) - 2L_k(\boldsymbol{\theta}_k^*)} \geq \mathfrak{z}_k^{\circ}(\beta)\right\}\right) \\ &\overset{\text{by (D.43)}}{\leq} I\!P\left(\bigcup\nolimits_{k=1}^{K}\left\{\sqrt{2L_k(\widetilde{\boldsymbol{\theta}}_k) - 2L_k(\boldsymbol{\theta}_k^*)} \geq \mathfrak{z}_k(\beta + \Delta_{\mathrm{full},\,k}) - \varepsilon_k\right\}\right) \\ &\overset{\text{by L. D.5 and (D.45)}}{\leq} \mathfrak{c}^{-1}(\beta + \Delta_{\mathrm{full,\,max}}) + \Delta_{\mathrm{ac,LR}}. \end{aligned}$$

Therefore, if $\mathfrak{c}(\alpha) \geq \Delta_{\mathrm{full,\,max}}$, then

$$I\!P\left(\bigcup\nolimits_{k=1}^{K}\left\{\sqrt{2L_k(\widetilde{\boldsymbol{\theta}}_k) - 2L_k(\boldsymbol{\theta}_k^*)} \geq \mathfrak{z}_k^{\circ}(\mathfrak{c}(\alpha) - \Delta_{\mathrm{full,\,max}})\right\}\right) \leq \alpha + \Delta_{\mathrm{ac,LR}}.$$

And by (D.47) for $\mathfrak{c}^{\circ}(\alpha) \geq 2\Delta_{\text{full, max}}$ it holds

$$\begin{aligned}
&I\!P\left(\bigcup\nolimits_{k=1}^{K}\left\{\sqrt{2L_k(\widetilde{\boldsymbol{\theta}}_k) - 2L_k(\boldsymbol{\theta}_k^*)} \geq \mathfrak{z}_k^{\circ}\left(\mathfrak{c}^{\circ}(\alpha) - 2\Delta_{\text{full, max}}\right)\right\}\right) - \alpha \\
&\leq \Delta_{\text{total}} + 2\Delta_{\text{ac,LR}}.
\end{aligned}$$

Similarly for the inverse direction:

$$\begin{aligned}
\mathfrak{c}^{\circ\,-1}(\beta) &\leq I\!P^{\circ}\left(\bigcup\nolimits_{k=1}^{K}\left\{\sqrt{2L_k^{\circ}(\widetilde{\boldsymbol{\theta}}_k^{\circ}) - 2L_k^{\circ}(\widetilde{\boldsymbol{\theta}}_k)} \geq \mathfrak{z}_k^{\circ}\left(\beta\right)\right\} - \varepsilon_{1,k}\right) \\
&\leq I\!P^{\circ}\left(\bigcup\nolimits_{k=1}^{K}\left\{\sqrt{2L_k^{\circ}(\widetilde{\boldsymbol{\theta}}_k^{\circ}) - 2L_k^{\circ}(\widetilde{\boldsymbol{\theta}}_k)} \geq \mathfrak{z}_k\left(\beta + \Delta_{\text{full},\,k}\right) - \varepsilon_{1,k} - \varepsilon_k\right\}\right) \\
&\leq I\!P\left(\bigcup\nolimits_{k=1}^{K}\left\{\sqrt{2L_k(\widetilde{\boldsymbol{\theta}}_k) - 2L_k(\boldsymbol{\theta}_k^*)} \geq \mathfrak{z}_k\left(\beta + \Delta_{\text{full},\,k}\right)\right\}\right) + \Delta_{\text{total}} + \Delta_{\text{ac,LR}} \\
&\leq \mathfrak{c}^{-1}(\beta + \Delta_{\text{full, max}}) + \Delta_{\text{total}} + \Delta_{\text{ac,LR}},
\end{aligned}$$

where $0 \leq \varepsilon_{1,k} \leq \mathtt{C}(p_k + \mathtt{x})/\sqrt{n}$. This implies

$$\begin{aligned}
\mathfrak{c}^{\circ\,-1}(\beta) &\leq \mathfrak{c}^{-1}(\beta + \Delta_{\text{full, max}}) + \Delta_{\text{total}} + \Delta_{\text{ac,LR}}, \\
\mathfrak{c}^{\circ}(\alpha) &\geq \mathfrak{c}\left(\alpha - \Delta_{\text{total}} - \Delta_{\text{ac,LR}}\right) - \Delta_{\text{full, max}}.
\end{aligned} \tag{D.48}$$

$$\begin{aligned}
&I\!P\left(\bigcup\nolimits_{k=1}^{K}\left\{\sqrt{2L_k(\widetilde{\boldsymbol{\theta}}_k) - 2L_k(\boldsymbol{\theta}_k^*)} \geq \mathfrak{z}_k^{\circ}(\beta + \Delta_{\text{full},\,k})\right\}\right) \\
&\overset{\text{by (D.43)}}{\geq} I\!P\left(\bigcup\nolimits_{k=1}^{K}\left\{\sqrt{2L_k(\widetilde{\boldsymbol{\theta}}_k) - 2L_k(\boldsymbol{\theta}_k^*)} \geq \mathfrak{z}_k\left(\beta\right)\right\}\right) \\
&\geq \mathfrak{c}^{-1}(\beta) - \Delta_{\text{ac,LR}}.
\end{aligned}$$

$$I\!P\left(\bigcup\nolimits_{k=1}^{K}\left\{\sqrt{2L_k(\widetilde{\boldsymbol{\theta}}_k) - 2L_k(\boldsymbol{\theta}_k^*)} \geq \mathfrak{z}_k^{\circ}(\mathfrak{c}(\alpha) + \Delta_{\text{full, max}})\right\}\right) \geq \alpha - \Delta_{\text{ac,LR}}.$$

And by (D.48)

$$\begin{aligned}
&I\!P\left(\bigcup\nolimits_{k=1}^{K}\left\{\sqrt{2L_k(\widetilde{\boldsymbol{\theta}}_k) - 2L_k(\boldsymbol{\theta}_k^*)} \geq \mathfrak{z}_k^{\circ}(\mathfrak{c}^{\circ}(\alpha) + 2\Delta_{\text{full, max}})\right\}\right) - \alpha \\
&\geq -\Delta_{\text{total}} - 2\Delta_{\text{ac,LR}}.
\end{aligned}$$

for

$$\Delta_{\mathfrak{z},\,\text{total}} \overset{\text{def}}{=} \Delta_{\text{total}} + 2\Delta_{\text{ac,LR}} \leq 3\Delta_{\text{total}}. \tag{D.49}$$

Conditions of Theorem 3.1 include $z_k \geq C\sqrt{p_k}$, therefore, it has to be checked that $\mathfrak{z}_k^{\circ}(\alpha) \geq C\sqrt{p_k}$. It holds by Theorem A.4, Proposition C.1, Lemmas A.2 and D.2 with probability $\geq 1 - 12\mathrm{e}^{-\mathtt{x}}$:

$$\begin{aligned}
&I\!P^{\circ}\left(\sqrt{2\{L_k^{\circ}(\widetilde{\boldsymbol{\theta}}_k^{\circ}) - L_k^{\circ}(\widetilde{\boldsymbol{\theta}}_k)\}} > \mathtt{C}\sqrt{p_k - \sqrt{2\mathtt{x}p_k}} + \mathtt{C}(p_k + \mathtt{x})/\sqrt{n}\right) \\
&\qquad \geq 1 - 8\mathrm{e}^{-\mathtt{x}},
\end{aligned}$$

Taking $1 - 8\mathrm{e}^{-\mathtt{x}} \geq \alpha$, we have

$$\mathfrak{z}_k^{\circ}(\alpha) \geq \mathtt{C}\sqrt{p_k - \sqrt{2\mathtt{x}p_k}} + \mathtt{C}2(p_k + \mathtt{x})/\sqrt{n}.$$

Inequalities for $\mathfrak{c}^{\circ}(\alpha)$ had been already derived in (D.47) and (D.48) with

$$\Delta_{\mathfrak{c}} \stackrel{\text{def}}{=} \Delta_{\text{total}} + \Delta_{\text{ac,LR}}. \tag{D.50}$$

□

Lemma D.5. *Let the conditions from Section 3.4.1 be fulfilled, and the values $z_k \geq \sqrt{p_k}$ and $\delta_{z_k} \geq 0$ be s.t. $\mathtt{C} \max\limits_{1\leq k\leq K}\{n^{-1/2}, \delta_{z_k}\} \leq \Delta_{\varepsilon} \leq \mathtt{C} \min\limits_{1\leq k\leq K}\{1/z_k\}$ (Δ_{ε} is given in* (C.3)*), then it holds with probability $\geq 1 - 12\mathrm{e}^{-\mathtt{x}}$*

$$\begin{aligned}
&I\!P\left(\bigcup\nolimits_{k=1}^{K}\Big\{\sqrt{2L_k(\widetilde{\boldsymbol{\theta}}_k) - 2L_k(\boldsymbol{\theta}_k^*)} \geq z_k\Big\}\right) \\
&\quad - I\!P\left(\bigcup\nolimits_{k=1}^{K}\Big\{\sqrt{2L_k(\widetilde{\boldsymbol{\theta}}_k) - 2L_k(\boldsymbol{\theta}_k^*)} \geq z_k + \delta_{z_k}\Big\}\right) \leq \Delta_{\text{ac,LR}},
\end{aligned}$$

where

$$\Delta_{\text{ac,LR}} \leq 12.5\mathtt{C}\left(\frac{p_{\max}^3}{n}\right)^{1/8} \log^{9/8}(K)\log^{3/8}(np_{\text{sum}})\widehat{\mathfrak{a}}^{3/4}.$$

Proof of Lemma D.5. This statement's proof is similar to the one of Theorem 3.1 (see Section D.2.2). Here instead of the bootstrap statistics we consider only the values from the $\boldsymbol{Y}$-world. Let us denote $\mathtt{x}_2 \stackrel{\text{def}}{=} \mathtt{x} + \log(K)$. It holds with probability $\geq 1 - 12\mathrm{e}^{-\mathtt{x}}$

$$\begin{aligned}
&I\!P\left(\bigcup\nolimits_{k=1}^{K}\Big\{\sqrt{2L_k(\widetilde{\boldsymbol{\theta}}_k) - 2L_k(\boldsymbol{\theta}_k^*)} > z_k\Big\}\right) \\
&\stackrel{\text{L. A.7}}{\leq} I\!P\left(\bigcup\nolimits_{k=1}^{K}\Big\{\|\boldsymbol{\xi}_k\| > z_k - \Delta_{\text{w},k}(\mathtt{r}_{0,k}, \mathtt{x}_2)\Big\}\right) \\
&\stackrel{\text{Pr. C.1}}{\leq} I\!P\left(\bigcup\nolimits_{k=1}^{K}\Big\{\|\boldsymbol{\xi}_k\| > z_k + \delta_{z_k} + \Delta_{\text{w},k}(\mathtt{r}_{0,k}, \mathtt{x}_2)\Big\}\right) + \Delta_{\text{ac,LR}} \\
&\leq I\!P\left(\bigcup\nolimits_{k=1}^{K}\Big\{\sqrt{2L_k(\widetilde{\boldsymbol{\theta}}_k) - 2L_k(\boldsymbol{\theta}_k^*)} > z_k + \delta_{z_k}\Big\}\right) + \Delta_{\text{ac,LR}}\,,
\end{aligned}$$

where

$$\Delta_{\text{ac,LR}} \leq 12.5\mathtt{C}\left(p_{\max}^3/n\right)^{1/8}\log^{9/8}(K)\log^{3/8}(np_{\text{sum}})\widehat{\mathfrak{a}}^{3/4}.$$

Similarly to (D.40) and (D.41) the term $\Delta_{\text{ac,LR}}$ is equal to Δ_{ℓ_2} from Proposition C.1 with $\Delta_{\Sigma}^2 := 0$, $\delta_{z_k} := \delta_{z_k} + 2\Delta_{\text{w},k}(\mathtt{r}_{0,k}, \mathtt{x} + \log(K))$. □

D.2.4 Proof of Theorem 3.3

Proof of Theorem 3.3. Let us denote $\mathtt{x}_2 \stackrel{\text{def}}{=} \mathtt{x} + \log(K)$. By Lemmas A.7, A.8 and D.4 it holds with probability $\geq 1 - 12\mathrm{e}^{-\mathtt{x}}$

$$
\begin{aligned}
&I\!P^{\circ}\left(\bigcup\nolimits_{k=1}^{K}\Big\{\sqrt{2L_k^{\circ}(\widetilde{\boldsymbol{\theta}}_k^{\circ}) - 2L_k^{\circ}(\widetilde{\boldsymbol{\theta}}_k)} > z_k\Big\}\right) \\
&\geq I\!P^{\circ}\left(\bigcup\nolimits_{k=1}^{K}\Big\{\|\boldsymbol{\xi}_k^{\circ}(\boldsymbol{\theta}_k^*)\| > z_k + \Delta_{\mathrm{w},k}^{\circ}(\mathtt{r}_{0,k},\mathtt{x}_2) + \Delta_{\xi,k}^{\circ}(\mathtt{r}_{0,k},\mathtt{x}_2)\Big\}\right) \\
&\geq I\!P\left(\bigcup\nolimits_{k=1}^{K}\Big\{\|\widetilde{\boldsymbol{\xi}}_k\| > z_k - \Delta_{\mathrm{w},k}(\mathtt{r}_{0,k},\mathtt{x}_2)\Big\}\right) - \Delta_{\mathrm{b,\,total}} \qquad \text{(D.51)} \\
&\geq I\!P\left(\bigcup\nolimits_{k=1}^{K}\Big\{\|\boldsymbol{\xi}_k\| > z_k - \Delta_{\mathrm{w},k}(\mathtt{r}_{0,k},\mathtt{x}_2)\Big\}\right) - \Delta_{\mathrm{b,\,total}} \qquad \text{(D.52)} \\
&\geq I\!P\left(\bigcup\nolimits_{k=1}^{K}\Big\{\sqrt{2L_k(\widetilde{\boldsymbol{\theta}}_k) - 2L_k(\boldsymbol{\theta}_k^*)} > z_k\Big\}\right) - \Delta_{\mathrm{b,\,total}},
\end{aligned}
$$

here $\widetilde{\boldsymbol{\xi}}_k \stackrel{\text{def}}{=} \big(D_k^{-1}H_k^2D_k^{-1}\big)^{1/2}(\operatorname{Var}\boldsymbol{\xi}_k)^{-1/2}\boldsymbol{\xi}_k$, and $\Delta_{\mathrm{b,total}}$ is given below. Using the same notations as in the proof of Lemma D.4, we have

$$
\begin{aligned}
\widetilde{\Xi} &\stackrel{\text{def}}{=} \big(\widetilde{\boldsymbol{\xi}}_1^{\top},\ldots,\widetilde{\boldsymbol{\xi}}_K^{\top}\big)^{\top} \\
&= \big(\widehat{D}^{-1}\widehat{H}^2\widehat{D}^{-1}\big)^{1/2}(\operatorname{Var}\Xi)^{-1/2}\Xi,
\end{aligned}
$$

and by Theorem D.2 and by conditions $(\boldsymbol{\mathcal{I}_k})$, $(\boldsymbol{\mathcal{I}_{B,k}})$, it holds with probability $\geq 1 - \mathrm{e}^{-\mathtt{x}}$

$$
\begin{aligned}
\big\|\operatorname{Var}\widetilde{\Xi} - \operatorname{Var}^{\circ}\Xi^{\circ}\big\|_{\max} &= \big\|\widehat{D}^{-1}\big(\widehat{H}^2 - \widehat{\mathcal{V}}^2\big)\widehat{D}^{-1}\big\|_{\max} \\
&\leq \delta_{\widehat{\mathcal{V}}}^2(\mathtt{x})\big\|\widehat{D}^{-1}\widehat{H}^2\widehat{D}^{-1}\big\| \\
&\leq \delta_{\widehat{\mathcal{V}}}^2(\mathtt{x})(\widehat{\mathfrak{a}}^2 + \widehat{\mathfrak{a}}_B^2).
\end{aligned}
$$

Thus, inequality (D.51) follows from Proposition C.1 applied to the sets of vectors $\boldsymbol{\xi}_1^{\circ}(\boldsymbol{\theta}_1^*),\ldots,\boldsymbol{\xi}_K^{\circ}(\boldsymbol{\theta}_K^*)$ and $\widetilde{\boldsymbol{\xi}}_1,\ldots,\widetilde{\boldsymbol{\xi}}_K$. The error term $\Delta_{\mathrm{b,total}}$ is equal to Δ_{total} from Theorem D.2.2 (see (D.40), (D.41)) with $\widehat{\delta}_{\mathrm{smb}}^2 := 0$, thus

$$
\Delta_{\mathrm{b,total}} \leq 25\mathtt{C}\left(\frac{p_{\max}^3}{n}\right)^{1/8}\log^{9/8}(K)\log^{3/8}(np_{\mathrm{sum}})\left\{\big(\widehat{\mathfrak{a}}^2 + \widehat{\mathfrak{a}}_B^2\big)\big(1 + \delta_{\widehat{\mathcal{V}}}^2(\mathtt{x})\big)\right\}^{3/8}.
$$

Inequality (D.52) is implied by definitions of $\widetilde{\boldsymbol{\xi}}_k$ and matrices H_k^2, V_k^2, indeed:

$$
\begin{aligned}
&\Big\|\big(D_k^{-1}H_k^2D_k^{-1}\big)^{-1/2}\operatorname{Var}\boldsymbol{\xi}_k\big(D_k^{-1}H_k^2D_k^{-1}\big)^{-1/2}\Big\| \\
&\leq \Big\|\big(D_k^{-1}H_k^2D_k^{-1}\big)^{1/2}\big(D_kH_k^{-2}V_k^2H_k^{-2}D_k\big)\big(D_k^{-1}H_k^2D_k^{-1}\big)^{1/2}\Big\| \\
&\leq 1,
\end{aligned}
$$

therefore, $\|\widetilde{\boldsymbol{\xi}}_k\| \geq \|\boldsymbol{\xi}_k\|$.

The second inequality in the statement is proven similarly to (D.47). It implies together with Theorem 2.4 the rest part of the statement, having

$$\Delta_{\mathrm{b},\mathfrak{c}} \stackrel{\mathrm{def}}{=} \Delta_{\mathrm{b},\,\mathrm{total}} + \Delta_{\mathrm{ac,LR}}. \tag{D.53}$$

□

Bibliography

Aerts, M. and Claeskens, G. (2001). Bootstrap tests for misspecified models, with application to clustered binary data. *Computational statistics & data analysis*, 36(3):383–401.

Arlot, S., Blanchard, G., and Roquain, E. (2010a). Some nonasymptotic results on resampling in high dimension. I. Confidence regions. *The Annals of Statistics*, 38(1):51–82.

Arlot, S., Blanchard, G., and Roquain, E. (2010b). Some nonasymptotic results on resampling in high dimension, II: Multiple tests. *The Annals of Statistics*, 38(1):83–99.

Barbe, P. and Bertail, P. (1995). *The weighted bootstrap*, volume 98. Springer.

Barsov, S. S. and Ul'yanov, V. V. (1987). Difference of Gaussian measures. *Journal of Soviet Mathematics*, 38(5):2191–2198.

Benjamini, Y. (2010). Simultaneous and selective inference: current successes and future challenges. *Biometrical Journal*, 52(6):708–721.

Bentkus, V. (2003). On the dependence of the Berry–Esseen bound on dimension. *Journal of Statistical Planning and Inference*, 113(2):385–402.

Beran, R. (1988). Balanced simultaneous confidence sets. *Journal of the American Statistical Association*, 83(403):679–686.

Beran, R. (1990). Refining bootstrap simultaneous confidence sets. *Journal of the American Statistical Association*, 85(410):417–426.

Berry, A. C. (1941). The accuracy of the Gaussian approximation to the sum of independent variates. *Transactions of the American Mathematical Society*, 49(1):122–136.

Bhattacharya, R. and Holmes, S. (2010). An exposition of Götze's estimation of the rate of convergence in the multivariate central limit theorem. *arXiv preprint arXiv:1003.4254*.

Bickel, P. J. and Rosenblatt, M. (1973). On some global measures of the deviations of density function estimates. *The Annals of Statistics*, pages 1071–1095.

Bonferroni, C. E. (1936). *Teoria statistica delle classi e calcolo delle probabilita*. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze.

Bücher, A. and Dette, H. (2013). Multiplier bootstrap of tail copulas with applications. *Bernoulli*, 19(5A):1655–1687.

Cao, H. and Kosorok, M. R. (2011). Simultaneous critical values for t-tests in very high dimensions. *Bernoulli*, 17(1):347–394.

Chatterjee, S. and Bose, A. (2005). Generalized bootstrap for estimating equations. *The Annals of Statistics*, 33(1):414–436.

Chen, L. H. and Fang, X. (2011). Multivariate normal approximation by Stein's method: The concentration inequality approach. *arXiv preprint arXiv:1111.4073*.

Chen, X. and Pouzo, D. (2009). Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals. *Journal of Econometrics*, 152(1):46–60.

Chen, X. and Pouzo, D. (2015). Sieve Wald and QLR inferences on semi/nonparametric conditional moment models. *Econometrica*, 83(3):1013–1079.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2013a). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics*, 41(6):2786–2819.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2013b). Supplement to "Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors".

Chernozhukov, V., Chetverikov, D., and Kato, K. (2014a). Anti-concentration and honest, adaptive confidence bands. *The Annals of Statistics*, 42(5):1787–1818.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2014b). Central limit theorems and bootstrap in high dimensions. *arXiv preprint arXiv:1412.3661*.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2014c). Comparison and anti-concentration bounds for maxima of Gaussian random vectors. *Probability Theory and Related Fields*, 162:47–70.

Claeskens, G. and Van Keilegom, I. (2003). Bootstrap confidence bands for regression curves and their derivatives. *Annals of Statistics*, pages 1852–1884.

Dickhaus, T. (2014). *Simultaneous Statistical Inference: With Applications in the Life Sciences*. Springer.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26.

Efron, B. and Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

Esseen, C.-G. (1942). *On the Liapounoff limit of error in the theory of probability*. Almqvist & Wiksell.

Esseen, C. G. (1956). A moment inequality with an application to the central limit theorem. *Scandinavian Actuarial Journal*, 1956(2):160–170.

Fisher, R. A. (1956). *Statistical methods and scientific inference*. Oliver & Boyd, Edinburgh.

Götze, F. (1991). On the rate of convergence in the multivariate CLT. *The Annals of Probability*, pages 724–739.

Götze, F. and Zaitsev, A. Y. (2014). Explicit rates of approximation in the CLT for quadratic forms. *The Annals of Probability*, 42(1):354–397.

Hall, A. R. (2005). *Generalized method of moments*. Oxford University Press Oxford.

Hall, P. (1991). On convergence rates of suprema. *Probability Theory and Related Fields*, 89(4):447–455.

Hall, P. (1992). *The bootstrap and Edgeworth expansion*. Springer.

Hall, P. (1993). On Edgeworth expansion and bootstrap confidence bands in nonparametric curve estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 291–304.

Hall, P. and Horowitz, J. (2013). A simple bootstrap method for constructing nonparametric confidence bands for functions. *The Annals of Statistics*, 41(4):1892–1921.

Hall, P. and Pittelkow, Y. (1990). Simultaneous bootstrap confidence bands in regression. *Journal of Statistical Computation and Simulation*, 37(1-2):99–113.

Härdle, W. (1989). Asymptotic maximal deviation of M-smoothers. *Journal of Multivariate Analysis*, 29(2):163–179.

Härdle, W. and Marron, J. (1991). Bootstrap simultaneous error bars for nonparametric regression. *The Annals of Statistics*, pages 778–796.

He, X. (1997). Quantile curves without crossing. *The American Statistician*, 51(2):186–192.

Horowitz, J. L. (2001). The bootstrap. *Handbook of econometrics*, 5:3159–3228.

Huber, P. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. Proc. 5th Berkeley Symp. Math. Stat. Probab., Univ. Calif. 1965/66, 1, 221-233 (1967).

Hudson, D. (1971). Interval estimation from the likelihood function. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 256–262.

Janssen, A. and Pauls, T. (2003). How do bootstrap and permutation tests work? *Annals of statistics*, pages 768–806.

Janssen, P. (1994). Weighted bootstrapping of U-statistics. *Journal of statistical planning and inference*, 38(1):31–41.

Johnston, G. J. (1982). Probabilities of maximal deviations for nonparametric regression function estimates. *Journal of Multivariate Analysis*, 12(3):402–414.

Kim, K. I. and van de Wiel, M. A. (2008). Effects of dependence in high-dimensional multiple testing problems. *BMC bioinformatics*, 9(1):114.

Kline, P. and Santos, A. (2012). Higher order properties of the wild bootstrap under misspecification. *Journal of Econometrics*, 171(1):54–70.

Koenker, R. (2005). *Quantile regression*. Number 38. Cambridge university press.

Koenker, R. and Bassett Jr, G. (1978). Regression quantiles. *Econometrica: Journal of the Econometric Society*, pages 33–50.

Lavergne, P. and Patilea, V. (2013). Smooth minimum distance estimation and testing with conditional estimating equations: uniform in bandwidth theory. *Journal of Econometrics*, 177(1):47–59.

Leek, J. T. and Storey, J. D. (2008). A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences*, 105(48):18718–18723.

Lindeberg, J. W. (1922). Eine neue Herleitung des Exponentialgesetzes in der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 15(1):211–225.

Liu, R. Y. (1988). Bootstrap procedures under some non-i.i.d. models. *Ann. Statist.*, 16(4):1696–1708.

Liu, W. (2010). *Simultaneous inference in regression*. CRC Press.

Liu, Y. and Wu, Y. (2011). Simultaneous multiple non-crossing quantile regression estimation using kernel constraints. *Journal of nonparametric statistics*, 23(2):415–437.

Ma, S. and Kosorok, M. R. (2005). Robust semiparametric M-estimation and the weighted bootstrap. *Journal of Multivariate Analysis*, 96(1):190–217.

Mammen, E. (1992). *When does bootstrap work?*, volume 77. Springer.

Mammen, E. (1993). Bootstrap and wild bootstrap for high dimensional linear models. *The Annals of Statistics*, pages 255–285.

Manly, B. F. (2006). *Randomization, bootstrap and Monte Carlo methods in biology*, volume 70. CRC Press.

Miller, R. G. (1981). *Simultaneous statistical inference*. Springer.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384.

Neumann, M. H. and Polzehl, J. (1998). Simultaneous bootstrap confidence bands in nonparametric regression. *Journal of Nonparametric Statistics*, 9(4):307–333.

Newton, M. A. and Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 3–48.

Prokhorov, Y. V. and Ulyanov, V. V. (2013). Some approximation problems in statistics and probability. In *Limit Theorems in Probability, Statistics and Number Theory*, pages 235–249. Springer.

Qu, Z. (2008). Testing for structural change in regression quantiles. *Journal of Econometrics*, 146(1):170–184.

Röllin, A. (2013). Stein's method in high dimensions with applications. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 49(2):529–549.

Romano, J. P. and Wolf, M. (2005). Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association*, 100(469):94–108.

Shao, J. and Tu, D. (1995). *The jackknife and bootstrap*. Springer.

Shevtsova, I. (2010). An improvement of convergence rate estimates in the Lyapunov theorem. In *Doklady Mathematics*, volume 82, pages 862–864. Springer.

Šidák, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62(318):626–633.

Slepian, D. (1962). The one-sided barrier problem for Gaussian noise. *Bell System Technical Journal*, 41(2):463–501.

Spokoiny, V. (2012a). Parametric estimation. Finite sample theory. *The Annals of Statistics*, 40(6):2877–2909.

Spokoiny, V. (2012b). Supplement to "Parametric estimation. Finite sample theory".

Spokoiny, V. (2013). Bernstein-von Mises Theorem for growing parameter dimension. *arXiv preprint arXiv:1302.3430*.

Spokoiny, V. and Willrich, N. (2015). Bootstrap tuning in ordered model selection. *arxiv preprint arxiv:1507.05034*.

Spokoiny, V. and Zhilova, M. (2015). Bootstrap confidence sets under model misspecification. *Ann. Statist.*, 43(6):2653–2675.

Spokoiny, V. G. and Zhilova, M. M. (2013). Uniform properties of the local maximum likelihood estimate. *Automation and Remote Control*, 74(10):1656–1669.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, pages 1135–1151.

Talagrand, M. (2003). *Spin glasses: a challenge for mathematicians: cavity and mean field models*, volume 46. Springer.

Tropp, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434.

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer, New York.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical processes*. Springer, New York.

Wasserman, L. (2006). *All of nonparametric statistics*. Springer Science & Business Media.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 359–372.

Westfall, P. H. (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment*, volume 279. John Wiley & Sons.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the Econometric Society*, pages 1–25.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62.

Working, H. and Hotelling, H. (1929). Applications of the theory of error to the interpretation of trends. *Journal of the American Statistical Association*, 24(165A):73–85.

Wu, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics*, 14(4):1261–1295+.

Zhilova, M. (2015). Simultaneous likelihood-based bootstrap confidence sets for a large number of models. *arXiv preprint arXiv:1506.05779*.

List of Figures

List of Tables

Selbständigkeitserklärung

Ich erkläre, dass ich die vorliegende Arbeit selbständig und nur unter Verwendung der angegebenen Literatur und Hilfsmittel angefertigt habe.

Berlin, den 23.06.2015 Mayya Zhilova